2020-09

MONTE CARLO SIMULATION WITH CENSORED SAMPLING

Akin, Ezra W.

Monterey, CA; Naval Postgraduate School

http://hdl.handle.net/10945/66062

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

DISSERTATION

**MONTE CARLO SIMULATION
WITH CENSORED SAMPLING**

by

Ezra W. Akin

September 2020

Dissertation Supervisor: Roberto Szechtman

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2020	**3. REPORT TYPE AND DATES COVERED** Dissertation

4. TITLE AND SUBTITLE MONTE CARLO SIMULATION WITH CENSORED SAMPLING	**5. FUNDING NUMBERS**
6. AUTHOR(S) Ezra W. Akin	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

We consider Monte Carlo simulation in a setting where the samples are subject to random censoring. Such censoring occurs in settings as varied and diverse as perimeter protection, survival analysis, and electro-magnetic spectrum monitoring. We introduce and analyze two estimators: one based on empirical likelihood methods and another rooted in control variates ideas. We show that the proposed estimators can dramatically reduce the estimator variance in relation to the crude Monte Carlo estimator while not sacrificing computational speed.

14. SUBJECT TERMS Monte Carlo, random censoring, control variates, maximum likelihood, simulation, stratification	**15. NUMBER OF PAGES** 85
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

MONTE CARLO SIMULATION WITH CENSORED SAMPLING

Ezra W. Akin
Major, United States Marine Corps
BS, U.S. Naval Academy, 2009
MS, Operations Research, Naval Postgraduate School, 2017

Submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2020

Approved by:

Roberto Szechtman
Department of
Operations Research
Dissertation Supervisor

Michael P. Atkinson
Department of
Operations Research

Moshe Kress
Department of
Operations Research

Lucas C. Wilcox
Department of
Applied Mathematics

Kevin Glazebrook,
Management Science
Lancaster University

Roberto Szechtman
Department of
Operations Research
Dissertation Chair

Approved by:

W. Matthew Carlyle
Chair, Department of Operations Research

Orrin D. Moses
Vice Provost of Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

We consider Monte Carlo simulation in a setting where the samples are subject to random censoring. Such censoring occurs in settings as varied and diverse as perimeter protection, survival analysis, and electro-magnetic spectrum monitoring. We introduce and analyze two estimators: one based on empirical likelihood methods and another rooted in control variates ideas. We show that the proposed estimators can dramatically reduce the estimator variance in relation to the crude Monte Carlo estimator while not sacrificing computational speed.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

This work is motivated by a security force tasked with defending a perimeter or border against infiltration attempts by an attacking force. This perimeter is broken into sections based upon the terrain and abilities of the defenders to observe each section. The infiltrators behave probabilistically in that each time period they attempt to infiltrate across the perimeter based upon some probability distribution. The security force attempts to learn certain properties of this unknown probability distribution that governs or is underlying the infiltration attempts or attackers' behavior. At each time period, the defenders send out watchers to different sections of the perimeter but are constrained in that they cannot have a watcher on each section of the perimeter in each time period. After attempting to observe a number of infiltration attempts, the security forces seek to estimate the central tendency (the mean of the underlying probability distribution) of the infiltration attempts. When an infiltration attempt occurs on a section with a watcher or defender, that attempt is fully observed. However, when the infiltration attempt occurs along an unobserved section (no watcher assigned) that infiltration attempt is censored or unobserved and the security force only knows that that infiltration occurred along an unobserved section of the perimeter. Therefore, the security force is faced with the problem of estimation using both censored and uncensored observations of the infiltrations. Such censoring occurs in settings as varied and diverse as perimeter protection (as described here), survival analysis (time to failure of some entity or object, e.g., aircraft part), and electromagnetic spectrum monitoring.

More generally, in typical Monte Carlo simulation, independent and identically distributed samples are randomly drawn from a random variable whose mean needs to be computed, and the sample average serves as the natural estimator for the (said) mean. In this thesis we consider the problem of Monte Carlo simulation when the samples are subject to random censoring. In the censoring setting we consider, one of two things happen for each random sample drawn from the target distribution: Its value is observed (as in standard Monte Carlo), or the analyst can only conclude that the sample value lies within a subset of the sample space. The censoring is random

because the latter subset is random, drawn from a distribution with finite support.

In such a random censoring setting, we introduce and analyze two estimators: one based on empirical likelihood methods, and another rooted in control variates ideas. We show that the proposed estimators can dramatically reduce the estimator variance in relation to the crude Monte Carlo estimator while not sacrificing computational speed, and also establish a deep connection between the empirical likelihood and control variates estimators. From the operational viewpoint, this means that for significantly fewer observations or samples, an analyst (the security forces in the former example) can achieve the same level of confidence in their estimate of the mean of the unknown underlying probability distribution (that which guides the infiltration attempts in the example above). In fact, in some of the specific settings tested through simulation in this dissertation, the proposed estimators reduced the variance from the basic Monte Carlo estimator by up to 99 percent.

Not only does this result in much more rapid estimation, for a fixed level of confidence, but the proposed control variate estimator has a significantly reduced computational complexity. This means that it can be employed on devices with far less computational power (less power and less hardware) without sacrificing the speed of computing the estimate which is key to enabling edge-computing.

Acknowledgments

I would first like to acknowledge my dependence upon my Heavenly Father and thank Him for His constant love, provision, and guidance. Soli Deo Gloria.

I would also like to express the sincerest gratitude to my parents, especially my mother, for inculcating me with a love of learning and sense of duty to country.

A special thanks to my best friend and wife, Abby, for her boundless patience, constant love, and heartfelt support during this journey. I would also like to thank my wonderful children who reminded me that a world exists outside the bounds of research and dissertation writing!

Finally, I must express the deepest appreciation and indebtedness to the faculty of the Operations Research Department—particularly my advisor, Roberto Szechtman, and my PhD committee members: Michael Atkinson, Moshe Kress, Lucas Wilcox, and Kevin Glazebrook. Your deep knowledge, patience, humility, assistance, and guidance were crucial in this endeavor, and I will value your friendship the rest of my life. Thank you very much!

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1: Introduction

This chapter provides an overview of the problem we are solving, why the problem matters, current methods for solving it, our novel approach, some measures of how successful our method is compared to some current baseline methods, and finally, the limitations (constraints) and assumptions made for our approach. This research primarily fills the gap between the methods of control variates and maximum likelihood in the case of censoring. While the use of maximum likelihood estimation and empirical likelihood estimation have been deeply studied in many censoring settings, the use of control variates with censoring has not been as thoroughly developed. Further, the connection between control variates and maximum likelihood, in the censoring setting, is novel to the author's knowledge.

1.1 Problem Overview and Motivation

The following setting serves as the primary motivation for this dissertation.

1.1.1 Perimeter Protection Problem (simple example)

Consider a security force tasked with defending a perimeter or border against infiltration attempts by an attacking force. This perimeter is broken into sections based upon the terrain and abilities of the defenders to observe each section. The infiltrators behave randomly in that each time period they attempt to infiltrate across the perimeter based upon some probability distribution. The security force attempts to learn certain properties of this unknown probability distribution that governs or is underlying the infiltration attempts or attackers' behavior. At each time period, the defenders send out watchers to different sections of the perimeter but are constrained in that they cannot have a watcher on each perimeter section each time period. The case where the defenders can only send out a single watcher is termed the Singleton Setting whereas the case where the defenders can send out either a single watcher or multiple watchers (just less than the total number of sections) is termed the Combinatorial Setting. After attempting to observe a certain number of infiltration attempts, the security forces seek to estimate the central tendency (the mean of the underlying probability distribution) of the infiltration attempts. When an infiltration attempt occurs on a section with a watcher or defender, that attempt is fully observed. But, when the infiltration attempt occurs along an unobserved section (no watcher assigned there) that infiltration attempt is censored or unobserved and the security force only knows that that infiltration occurred along an unobserved section of the perimeter. Therefore, the security force is faced with the problem of estimation using both censored and uncensored observations of the infiltrations.

1.1.2 Motivation

This section now examines a few different application settings which motivate this research which are related to the problem described in the previous sections. While not comprehensive, these examples demonstrate the wide variety of real world situations which can be formulated similar to the problem defined above.

Perimeter Protection (expanded with notation)

Consider the above base perimeter problem where a perimeter must be defended by security forces using searchers, defenders, or watchers against non-strategic attackers, smugglers, or infiltrators. The perimeter can be any shape (roughly straight might represent a border, while roughly oval, circular, or rectangular might represent a city, base, or building, respectively). This perimeter being a (topologically) linear and continuous segment such that it maps to a bounded interval of the real line. In this scenario, the “time” epochs represent discrete or individual arrivals of the attackers at the perimeter. In other words, only one attacker arrives per time period t (although the “time” periods may differ in length of actual time) for $t \in [T]$ with the notation $[T]$ denoting the set of integers $\{1, 2, \ldots, T\}$. Each attacker attempts to cross the perimeter in a single location, represented by the sample Z_t from the random variable Z with sample space Ω.

The perimeter, when mapped to the real number line starting at zero, is the support Ω. For example, if the perimeter has a length of 20 kilometers, then the support for Z would be the interval $[0, 20]$, where the value of Z_t would represent the number of kilometers from the “start” of the interval to the attempted crossing location for the t’th infiltration. The perimeter is partitioned into integer m distinct sections or cells.

The set of all cells is denoted by $\mathcal{C} = \{c_1, \ldots, c_i, \ldots, c_m\}$. We assume that the points where the smugglers attempt to cross the perimeter follow an unknown probability distribution, namely Z.

The defenders use a random placement method for the sensors or watchers at the perimeter. The subset of sections being observed in time period t is denoted by $B_t \subseteq \mathcal{C}$. These sensors or watchers will detect the attackers crossing the perimeter given they are within detection range (crossing attempt is within that cell), in notation if $Z_t \in c_i$ then the crossing is detected if $c_i \in B_t$. The probability of a sensor or watcher being placed in cell c_i being q_i. But, each sensor or watcher can only detect across an interval or section (detection range) of the perimeter. If the attacker attempts to cross an observed portion of the perimeter, (say $Z_t \in c_i$ and $c_i in B_t$) then that crossing location (Z_t) is observed. But, on the other hand, if the attacker attempts to cross an unobserved portion of the perimeter, (say $Z_t \in c_i$ but $c_i \notin B_t$) then that crossing location (Z_t) is unobserved or censored. The goal of the defender is to estimate or infer the mean, μ of Z such that they can ultimately place more sensors or defenders in that cell or perimeter interval. Of note, it is natural for this type of problem to have some form of constraint on how many cells can be observed each time period.

It is worth noting here, although we do not further analyze or discuss it in the rest of this dissertation, that if Z is multi-modal an estimate of μ may be rather useless to the defender. Imagine that $Z \approx N(4,1) + N(16,1)$ in the above example of a 20 km perimeter. In this case $\mu = E[Z] = 10$. But, this portion (around the 10 km mark) of the perimeter is a horrible section to defend. One very beneficial aspect of the estimators developed in this dissertation is that they form estimates for the p_i's ($p_i = P(Z \in c_i)$) as part of the method for estimating the mean. This means that the defender (in this example) also gains a sense of the modes (in a multi-modal Z) as well as the spread and skewness of Z even though those moments of Z are never explicitly estimated.

Electro-Magnetic Spectrum

Now, imagine that Z, instead of representing crossing locations along a perimeter, represents one aspect of an electro-magnetic transmission and that a sensor is used each time period to monitor a specific band of that spectrum. This interval or set of intervals (for multiple sensors) being the cells or strata $c_i \in B_t$. The "listener," in this case, places sensors on the spectrum following a probabilistic method (possibly to make it more difficult to avoid) with the probability of a sensor being placed on strata c_i being the probability q_i. Therefore, during a time "epoch," if the transmission occurs in the cell with the sensor then the transmission is observed or sensed ($Z_t \in B_t$) and therefore an uncensored sample Z_t is gained by the "listener". If, on the other hand, the transmission occurs outside of that strata or set of strata ($Z_t \notin B_t$), then it is a censored sample and the "listener" only knows that the transmission occurred outside of the intervals being "sensed" ($Z_t \notin B_t$).

Survival Analysis

Now, imagine that Z represents the lifetimes (time till failure or death) associated with some set of entities (each entity being $t \in [T]$) and each strata (c_i) represents a possible examination period. The combinatorial problem represents the situation where multiple examinations can occur during the lifetime of the entity ($|B_t| > 1$) while the non-combinatorial or singleton problem, represents a situation where only a single portion (interval) of examination can occur ($|B_t| = 1$). Following the same pattern as the previous examples, if a failure or end-of-life event (the event that the examination seeks to discover, Z_t) occurs within one of the examination periods ($Z_t \in B_t$) then it is observed, otherwise it is censored and the "examiner" only knows that the event did not occur during the examination periods ($Z_t \notin B_t$). In this setting, the probability of the event occurring in examination period c_i is defined as $p_i = P(Z \in c_i)$. We also assume that the "examiner" leverages a random system to determine when the examination periods will occur, i.e., the probability that an examination will occur in the lifetime interval c_i being $q_i = P(c_i \in B_t)$. Further, the entire possible lifetime of the entity being examined is partitioned into the strata $c_i \in \mathcal{C}$. For a further look at the different specific situations that can be modeled by this type of formulation, see Chapter 2 and the section on Survival Analysis.

1.2 Current Methods and Proposed Solution

A current solution to the problem described in Section 1.1, or the simplest method of estimating the mean, is to use the ratio of times that $Z_t \in c_i$ given that $c_i \in B_t$, over the total number of times $c_i \in B_t$, this ratio being defined as $\bar{p}_{iT}$ (the standard comma in the subscript is dropped unless absolutely necessary). Combining these estimates for the probability of Z falling within a specific strata, with an in-strata sample mean of the uncensored observations, results in what we term the Censored Naive (CN) estimator. The issue with this approach is two-fold. First, a censored observation tells the analyst something about the rest of the strata, not just the one selected for observation (as will be examined in Chapter 4, this can result in unnecessarily high variance in the estimator). Second, this estimator is unstable (specifically, if the underlying magnitude of the unknown mean $\mu = E[Z]$ increases, the variance of the estimator also increases). This last compounds the first issue (these are more fully examined both in Chapter 3 and Chapter 4).

A second "current" solution to this problem is that of employing the method of Empirical Likelihood or Maximum Likelihood (the use of these is not novel in this setting as they are examined by Owen (2001) among others). This approach is fully detailed in Chapter 3 and avoids the two issues that plague the censored naive estimator. But, these improvements come at the price of a significantly increased computational cost to form the estimator from a set of realizations. Specifically, the Maximum Likelihood estimator developed in Chapter 3 requires the solution of a root equation using some form of optimization software possibly running Newton's or the Bisection method. While the purpose of this dissertation is not to conduct an in-depth computational analysis, it is well known that finding the root of a function can be computationally expensive (the actual computational cost depends highly on the precision desired, the specific method used, and the starting location or interval for that method). More on this discussion in Chapters 3 and 4.

As mentioned by Glasserman and Yu (2005), a key foundational principle of stochastic simulation is that a simulation estimate's accuracy can be significantly improved by leveraging known properties of the physical system or simulated model representing that physical system. In the problem examined in this dissertation, the primary known property leveraged is that the underlying distribution being estimated is a proper probability distribution, in that its probability mass or density is 1.

One well known and widely used method for variance reduction, in the simulation field, is the method of control variates. Additionally, one can use the also well known method of maximum likelihood estimation that has very nice guarantees (see Chapter 2 for more details on the history of these methods as well as Chapter 3 for more details into our use of the methods). The problem examined by this dissertation leverages both of these methods and finds a deep asymptotic connection between them which in this censoring setting is novel.

1.3 Measures of Success and Limitations

This section serves to list the ground rules used in this dissertation and is broken down into what measures of success we use (meaning how well or successful are the proposed estimators and how do we measure that performance) as well as what limitations or assumptions are made. The limitations will be broken down by topic, namely, the censoring process, the stratification or partitioning of the support for the random variable being observed Z, and the sampling process of Z. The following are the highlights, more detail is given in Chapters 3 and 4.

1.3.1 Measures of Success

We use two different estimators as a baseline for measuring the success of our proposed estimators. Specifically, these two baseline estimators are essentially the worst and best case scenarios. The Censored Naive (CN) estimator described in Section 1.2 serves as the worst case, i.e., our proposed estimators must outperform (by some measure) the CN estimator. The best case is defined as the situation where the analyst is always able to observe the value of Z_t, resulting in what we call the “uncensored” estimator, which serves as the maximum performance possible by any estimator operating with some form of censoring.

To measure the estimators’ performance, we primarily compute their variance, asymptotically, as T (the number of die rolls) grows. We also use the Mean Squared Error of each of the estimators compared against the true underlying Z distribution properties (known by us for the purposes of simulation but generally unknown to the analyst).

These measures are employed while varying different aspects of the censoring scheme and the underlying properties of Z. The results of these measures, applied to the different estimators, can be seen in Chapter 4.

Limitations: Censoring Process

This dissertation does not examine the effect of intelligently controlling the censoring scheme or adjusting its probabilities during the simulation process. We believe that adjusting the number of "looks" per strata may result in a more efficient estimate but leave this question for future work. Further, we assume that the censoring process is probabilistic (multinomial) and specifically that the probability that cell c_i is contained in B_t is q_i and that the cells c_i and c_j are contained together in B_t is q_{ij}. Additionally, $0 < q_i \leq 1$ and $0 \leq q_{ij} \leq 1$ which implies that for sufficiently large T, no matter how small the q_i, strata c_i will eventually be "looked" at or observed. Further, it is assumed that B_t is independent of both Z_t and all $B_1, \ldots, B_{t-1}$.

Limitations: Sampling and Stratification Processes

This dissertation assumes that each sample Z_t from Z is independent of both B_t and $Z_1, \ldots, Z_{t-1}$. It also assumes a finite partitioning of the support for Z, namely that there are m total strata and that these strata fully cover that support, i.e., $0 < p_i, \forall i \in [m]$ and that $\sum_{i=1}^{m} p_i = 1$. Further, we assume that there is no overlap between strata, i.e., $c_i \cap c_j = \emptyset$.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Literature Review

This work follows a few different strands of study, namely Point Estimation from the field of Statistical Inference with random interval Censoring of the samples using the methods of Maximum Likelihood Estimation and Control Variates. Therefore, the following is a brief summary of these research threads and some notable contributions to and advancements in these bodies of knowledge.

2.1 Statistical Inference

Statistical Inference is a field of statistics interested in inferring or deducing properties about an underlying probability distribution from a set of data or samples from that distribution. This data is assumed to be from a homogeneous population and therefore inference is possible. The set of assumptions made is called the statistical model. As David Cox (2006) noted: "How translation from subject-matter problem to statistical model is done is often the most critical part of an analysis." The specific area of statistical inference that this dissertation will focus on is that termed or called "point estimation."

The basic purpose of point estimation is to calculate a "best guess" or "best estimate" for a single value, usually a population parameter such as the mean, from a set of observations from a given population. Numerous estimators have been proposed with different underlying assumptions or requirements and different settings in which they tend to perform better or worse. Two primary methods are: the Method of Moments (MoM) and Maximum Likelihood Estimation (MLE). This dissertation will focus in part on developing a Maximum Likelihood Estimator and therefore this literature review will focus on the developments of that estimation technique. For a summary of recent developments in the Method of Moments applied to financial settings see Zivot and Wang (2007).

Maximum Likelihood Estimation

Edgeworth (1908*a*) (continued in Edgeworth (1908*b*)) notes that Maximum Likelihood Estimation has been used by such notable figures as Carl Friedrich Gauss, Pierre-Simon Laplace, and Thorvald N. Thiele. According to Aldrich (1997) and Pfanzagl (2011) its widespread use and popularity are primarily due to its use by Ronald Fisher and his analysis of it, although he was unable to obtain a proof. It wasn't until Samuel S. Willks obtained a proof in Wilks (1938), called Wilk's Theorem, that the Maximum Likelihood Estimation method had proven asymptotic guarantees. Wilks (1938) demonstrates that, asymptotically, the error resulting from estimation of the logarithm of likelihood values for independent observations is χ^2 in distribution. Maximum Likelihood Estimators can be roughly divided into two major types, those that assume a known underlying distribution for the data (parametric) and those that do not make that assumption (nonparametric).

The basic Maximum Likelihood Estimator is defined here in Table 2.1, this table is copied directly from Akin (2017) and Devore (2015).

Definition: Maximum Likelihood Estimator

Let $X_1, X_2, \ldots, X_n$ have a joint probability mass function or PDF of

$$f(x_1, x_2, \ldots, x_n \mid \theta_1, \theta_2, \ldots, \theta_m) \tag{2.1}$$

where the parameters $\theta_1, \theta_2, \ldots, \theta_m$ have unkown values. When $x_1, x_2, \ldots, x_n$ are the observed sample values and (2.1) is regarded as a function of $\theta_1, \theta_2, \ldots, \theta_m$, it is called the **likelihood function**. The maximum likelihood estimates (MLE's) $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ are those values of the θ_i's that maximize the likelihood function, so that

$$f(x_1, \ldots, x_n \mid \hat{\theta}_1, \ldots, \hat{\theta}_m) \;\geq\; f(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_m) \;\text{ for all }\; \theta_1, \ldots, \theta_m \tag{2.2}$$

When the X_i's are substituted in place of the x_i's, the **maximum likelihood estimators** result.

Table 2.1. Definition: Maximum Likelihood Estimator. Reproduced from Devore's *Probability and Statistics for Engineering and the Sciences*, see Devore (2015)

The likelihood function defined in Table 2.1 enables the calculation of a likelihood ratio based on the results of Wilk's Theorem. This enables the statistician to create confidence intervals and more specifically to conduct hypothesis testing on the maximum likelihood estimates. By the Neyman-Pearson lemma in Neyman and Pearson (1933), this test is the most powerful statistical test. Of note, while R.A. Fisher first coined the terms "efficiency," "sufficiency," and "consistency" in relation to the method of maximum likelihood estimation, it wasn't until Lehmann and Casella (1998) that this method had the following properties proven. Namely, these properties are, directly copied from Akin (2017):

1. **Asymptotic Consistency**
   MLEs converge in probability to the true parameter value, θ. Further, by increasing our sample size, n, we can also achieve an arbitrary level of precision, see Lehmann and Casella (1998).
2. **Asymptotic Efficiency**
   This means that as the sample size n increases (tends towards infinity) the MLE converges to the true parameter value, θ, as fast as the quickest possible method. In other words, this method converges as quickly as theoretically possible. It achieves the so-called Cramér-Rao lower bound, which means that no consistent estimator can converge more quickly, see Lehmann and Casella (1998); Cramér (2016); Rao (1992). In other words, while other consistent estimators may match an MLE in convergence rate, they are not able to beat it.
3. **Asymptotic Normality**
   Again, as n increases, the MLEs converge in distribution to a Gaussian (normal) distribution with the mean being equal to the true parameter value, θ, and a minimal variance, see Lehmann and Casella (1998). Which, according to Devore (2015) is *"...as small as or nearly as small as can be achieved by any estimator."*

One issue that arises though, is what to do when you want to assume nothing about the underlying distribution of the population, specifically, the situation of infinite dimensional statistical models. Murphy et al. (1997) provide a very useful summary of the initial work conducted in the so called semiparametric or nonparametric maximum likelihood estimation (NPMLE) field where the asymptotic normality and efficiency were established. Their main contribution is to move towards providing a general likelihood ratio theory for semiparametric models. They conclude that while "inference for semiparametric models is a complex problem... likelihood ratio inference will, in general, be available." Art Owen (2001) further explores this in his seminal work Empirical Likelihood and the connections between nonparametric likelihood and empirical likelihood. Owen (2001) summarizes the difference between using parametric and nonparametric MLE very well. He states: "Either... can be best, depending on our goals and some assumptions on the data." He concludes: "If the underlying distribution does not follow the parametric one, then the NPMLE will ordinarily be better, at least for large enough n." Much additional work has been done in this area, for one example with a monotonicity constraint, see Banerjee et al. (2007).

2.2 Censoring

The concept of censoring, in the statistical sense, can be traced back at least to Bernoulli (1766) with further work by Pearson and Lee (1908) and Fisher (1931) the latter who examined distributions that were in some sense "truncated". Pearson and Fisher were interested in the truncation of a normal distribution or population. This study of truncation was further studied by Stevens (1937) and Hald (1949). The term *censoring*, though, was not used until Hald (1949). He wished to distinguish between the truncation that results in random samples being taken from an "*incomplete* normal distribution" as opposed to where the samples from a "*complete* normal distribution" are truncated. He therefore termed the first *truncation* or samples from a *truncated* population or distribution (hence *incomplete*) and the second he labeled *censoring* or *censored* samples (where the population or distribution is still *complete*). This work was followed by Cohen Jr. (1950) and Gupta (1952). The study of *censoring* was further sub-divided by Gupta (1952) when he distinguished between what he called "Type-I" and "Type-II" censoring. Specifically, he distinguished between censoring of samples based on their value (all observations above, below, or between truncation-points or thresholds are censored) and censoring based on say the $(n - k)$ largest or smallest samples of n total samples being censored. He defined these two cases as "Type-I" and "Type-II" respectively. Further, in Gupta (1952), he noted that Stevens (1937), Hald (1949), and Cohen Jr. (1950) all studied the Type-I censoring problem. Currently, these two terms, Type-I and Type-II censoring, commonly have different meanings and therefore, the terms “value-based” and “sampling-based” censoring will be used here instead. These terms are selected since Gupta’s “Type-I” censoring is primarily based upon the sample’s value, while his “Type-II” censoring is primarily based upon a sampling methodology (a pre-determined number being censored, as opposed to all samples with values, say, above a truncation-point or threshold being censored).

Survival Analysis

Another area of research that is very closely related to the censoring problem is Survival Analysis, Life Testing, Hazard Functions, or Reliability Theory. This field is primarily focused on studying the lifetime of some subject, entity, or object and how long until an event occurs such as a failure or death for that entity. To complicate matters, this field also uses the terms Type-I and Type-II censoring but to refer to specific methodologies for sampling or testing lifetimes which therefore naturally fall under what here is called *sampling-based* censoring. Specifically, these definitions of Type-I and Type-II censoring refer to a set number of entities being observed for their failure times. The difference being the stopping criteria for the experiment. In the Type-I case, the experiment is stopped at a pre-determined time, resulting in all entities without a failure being censored (specifically, their lifetimes are *right-censored*). In the Type-II case, the experiment is stopped after a pre-determined number of failures, resulting in the remaining entities being censored. Another version of censoring studied in this field is the so-called *random* censoring. This form is the result of the observation or censoring time for the entity being random and statistically independent of the actual failure time for that entity. Therefore, if the failure occurs after the censoring time then that sample is censored (an instance of *value-based* censoring). Additionally, in current terminology, *right*, *left*, *interval*, and *double* censoring all refer to what here is called *value-based* censoring. A quick summary or set of examples for these different types of censoring is given by Owen (2001) in his chapter on Biased and Incomplete Samples (Chapter 6).

Of note, many *value-based* censoring methods assume a fixed set of truncation points. The research presented in this work is conditionally *value-based* in that at each time period, an external censoring process determines a new set of truncation points for that time period. Then, if the sample falls within the defined interval(s) (or strata), it is uncensored, otherwise its exact value is censored.

In the Survival Analysis literature, the focus is primarily on the lifetime of an entity. Therefore, the underlying probability distributions being studied tend to be restricted to those that best represent a lifetime (e.g., Weibull and exponential). Specifically, those with continuous support on the positive reals. The methods presented in this work, in Chapter 3, can be used in this setting but are also applicable to both discrete and continuous distributions with finite or infinite support.

Another type of censoring in the Survival Analysis literature that is worth noting is *middle* or *interval* censoring. In this situation, data or samples become censored or unobservable when they fall within a censoring interval. Owen (2001) and Davarzani and Parsian (2011) provide good examples of this type of censoring. In notation, each of the n entities within the experiment have lifetimes $T_1, \dots, T_n$ and an independent but associated censoring interval $[L_1, R_1], \dots, [L_2, R_2]$. This results in the i'th entity's lifetime T_i being observable only if $T_i \notin [L_i, R_i]$. This form of censoring was examined by Owen (2001) as well as by Jammalamadaka and Mangalam (2003). Jammalamadaka and Mangalam (2003) obtained both a self-consistent estimator and a non-parametric maximum likelihood estimator for this censoring type. These were obtained for continuous lifetime type data while Davarzani and Parsian (2011) extended the analysis to discrete or count forms of lifetime data. For example, instead of length of time a copier has been running since its last servicing, maybe a more useful or analogous measure would be the number of copies made. This idea of *middle*-censoring provides a useful background to that examined in this dissertation.

The inverse to *middle*-censoring, that also arises in the Survival Analysis literature (very closely related to the censoring used in this dissertation), is called *double*-censoring or *doubly* censored data. This term refers to data that is both *right* and *left* censored. In essence, a failure is only observed when it occurs within a "window" or interval. Hence, this being the inverse of the *middle*-censoring case. An example of this in the literature is given by both Owen (2001) and by Chen and Zhou (2003). This form of censoring is formally defined in the following way. In notation, the random variable $Z_t \in \mathbb{R}$ is fully observed or uncensored when $Z_t \in [L_t, R_t]$, with $[L_t, R_t]$ being the censoring interval. This also means that Z_t is left censored to $(-\infty, L_t)$ if $Z_t < L_t$ but if $R_t < Z_t$ it is right censored to (R_t, ∞). Of note, in this definition $L_t < R_t$ and both may vary (deterministically or randomly) each time period $t \in 1, \dots, T$. This form of censoring is nearly identical to that used in this dissertation within the *singleton* or *non-combinatorial* setting since the interval being observed is contiguous. In our *combinatorial* setting, the strata being observed are not necessarily contiguous.

2.3 Nonparametric MLE with Censoring

The combination of the ideas of censored data with maximum likelihood estimation, and specifically, nonparametric likelihood methods, has been extensively studied. Owen (2001) examines this as part of his seminal work Empirical Likelihood. This work was further studied by Jammalamadaka and Mangalam (2003) in the continuous case and extended by Davarzani and Parsian (2011) to the discrete case for the so-called *middle*-censoring problem for lifetime data. But, Owen (2001) summarizes the key asymptotic results for the *double* censoring problem, as follows (copied from Chapter 6 in Owen (2001) and based on Murphy et al. (1997)).

Given that $(X_t, Y_t, Z_t) \in \mathbb{R}^3$ are all independent and identically distributed (with F_X, F_Y, F_Z being the distributions respectively) let X_t be *doubly* censored by Y_t on the right and by Z_t on the left. Hence, X_t is observed if and only if $Z_t \leq X_t \leq Y_t$ and we let $U_t = X_t$ and $\delta = 0$. Otherwise, if $Y_t < X_t$ then X_t is censored to (Y_t, ∞) and we let $U_t = Y_t$ and $\delta = 1$. But, if $X_t < Z_t$ then X_t is censored to $(-\infty, Z_t)$ and we let $U_t = Z_t$ and $\delta = -1$. Here $Z_t \leq Y_t$. With these defined, the conditional likelihood function for F_X is,

$$L_c(F) = \prod_{t=1}^{T} F\big(U_t\big)^{\delta_t = 0} F\big((U_t, \infty)\big)^{\delta_t = 1} F\big((-\infty, U_t)\big)^{\delta_t = -1},$$

with x^A being used as a shorthand for the indicator, x^{1_A}. Now, letting $\theta = \int q(x) dF_X(x)$ with θ_0 indicating realizations of X_t, for a known function q, we define

$$\mathcal{R}(\theta) = \frac{\max\{L_c(F) \mid \int q(x) dF_X(x) = \theta\}}{\max_F\{L_c(F)\}}.$$

These definitions lead to the following theorem (Theorem 6.10 in Owen (2001) with proof in Murphy et al. (1997)), copied directly from Owen (2001).

Theorem *Let $\mathcal{R}$ and θ be as described above. Suppose that F_X, F_Y, and F_Z are continuous distributions, that $F_X([A, B]) = 1$, for some $0 \leq A < B < \infty$, that $Pr(Z < u \leq Y) > 0$ for all $u \in [A, B]$, that $F_Z\big([0, B)\big) = 1$, and that $F_Y\big([0, A)\big) = 0$. Suppose that q is a left continuous function of bounded variation on $[A, B]$ with $\int q^2(x)dF_X(x) - \big(\int q(x)dF_X(x)\big)^2 > 0$. Then $-2\log\mathcal{R}(\theta_0) \to \chi^2_{(1)}$.*

This area was further studied by Chen and Zhou (2003) who obtained a constrained NPMLE for *doubly* censored data by extending the self-consistent equation given by Gill et al. (1989). Additional work was done by Zhou (2005) who proposed an algorithm to compute the empirical likelihood ratio for censored/truncated lifetime data under mean type constraints by modifying the self-consistent/Expectation-Maximizing (EM) algorithm. The majority of this work is related to Survival Analysis in that the underlying distributions are used primarily to model lifetimes.

One other interesting development is the work of Fygenson et al. (1994) who looked at using a stratification scheme to increase the efficiency of these estimators while also increasing the applicability to different censoring situations arising in Survival Analysis. While only looking at the *right* censoring case, they develop an "artificial stratification" scheme, an extension of the "synthetic data" method proposed by Koul et al. (1981) and improved by Leurgans (1987). This method of "artificial stratification" is as follows. Consider a right censoring lifetime analysis situation where the response to a linear regression model $Y_t = \boldsymbol{X}_t\boldsymbol{\beta}+\varepsilon_t$, is censored such that one observes the tuple $(U_t, \delta_t, \boldsymbol{X}_t)$ where,

$$U_t = \min(Y_t, C_t), \qquad \delta_t = I_{[Y_t \leq C_t]}, \qquad t = 1, 2, \ldots, T,$$

and where the C_t are independent random variables that are also independent of ε_t. Further, I_A is the indicator function of event A. The idea behind the "artificial stratification" scheme is that the C_t are independent and identically distributed within each of m strata allowing for group dependence of the censoring distribution on the covariates. This “stratification” is such that,

$$\begin{array}{rl} C_1, C_2, \ldots, C_{t_1} & \text{are i.i.d. as } G_1(\cdot); \\ C_{t_1+1}, C_{t_1+1}, \ldots, C_{t_2} & \text{are i.i.d. as } G_2(\cdot); \\ & \vdots \\ C_{t_1+\ldots+t_{m-1}+1}, \ldots, C_{t_1+\ldots+t_{m-1}+t_m} & \text{are i.i.d. as } G_m(\cdot). \end{array}$$

Therefore, instead of treating all of the C_t censoring times as i.i.d., they are treated as independent random variables from the $m > 1$ strata. The result of this method is an unbiased and more efficient estimator for $\boldsymbol{\beta}$ in linear regression models with random censoring (assuming censoring is of the *right* censoring form) given large samples, Fygenson et al. (1994). Ideas from this work are used in this dissertation but with a fundamentally different form of censoring. Namely, this dissertation is focused, in the singleton case, on something very similar to what Owen (2001) termed *doubly* censored data, while the form of censoring in our combinatorial case is to the author’s knowledge a novel form of censoring.

Of note, in many truncation and censoring settings asymptotic χ^2 distributions have been found to hold providing useful asymptotic results and specifically enabling the creation of confidence intervals and hypothesis testing. A list and associated summary, of some of the more important results, are provided in Chapter 6 of Owen (2001)

2.4 Control Variates

As mentioned by Glasserman and Yu (2005), a key foundational principle of stochastic simulation is that a simulation estimate’s accuracy can be significantly improved by leveraging known properties of the physical system or simulated model representing that physical system. One of the most powerful and widely known and used methods for this variance reduction, in the simulation field, is the method of control variates, as noted by Glynn and Szechtman (2002).

The basic simulation setting where Control Variates (CV) is used is as follows, summarized from Glynn and Szechtman (2002). Consider an analyst who is attempting to compute or estimate a quantity μ that can be expressed as the expectation of a random variable (rv) X with distribution F such that $\mu = E[X]$. The standard method for estimating this quantity, is to use the sample mean across n independent and identically distributed (iid) samples from F, denoted by $X_1, X_2, \ldots, X_n$. The sample mean is therefore, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. The method of control variates is based on the idea of selecting a random variable, Y with known mean ν that is correlated or jointly distributed with X. This correlation is then used to reduce the variance in the estimator. In notation, the control variate random variable $C = Y - \nu$ is guaranteed to have a mean of zero (by construction) and hence $X(\lambda) = X - \lambda C$ is an estimator for μ. Therefore, the analyst is able to compute a better (than the sample mean) estimate for μ by generating iid samples $(X_1, C_1), (X_2, C_2), \ldots, (X_n, C_n)$ of the pair (X, C) and forming the following estimate of μ via $\bar{X}_n(\lambda) = \bar{X}_n - \lambda\bar{C}_n$. In this, $\bar{C}_n$ and $\bar{X}_n$ are sample means respective to C and X and $\lambda \in \mathbb{R}$ is an arbitrary scalar. This control coefficient, λ, if chosen judiciously, can result in a variance reduction in the original estimator $\bar{X}_n$ resulting in a better estimate for μ. The choice of λ must therefore be based in such a way as to maximize the variance reduction–which is achieved when the analyst uses $\lambda^* = \frac{\text{Cov}(X,C)}{\text{Var}(C)}$.

For additional detail on this method, see Law (2007), Ross (1972), Glasserman and Yu (2005), and Ross (2014) as well as Chapter 3 where this will be further examined and used. For early examples of the use of control variates in the field of Econometrics see Appendix D in Sargan (1976) and for recent uses of control variates in importance sampling for Bayesian computation with regression models see Kharroubi (2018).

Of note, some important extensions of the method of control variates are as follows. Glynn and Szechtman (2002) examined the use of multiple control variates, essentially where the analyst has n different control variates each with their own mean. Further, they established connections between control variates and conditional Monte Carlo, antithetics, rotation sampling, stratification, and nonparametric maximum likelihood. This last being most closely related to the results developed within this dissertation (their work is in the uncensored setting whereas this dissertation looks at the censored setting). One key result though, is that under weak assumptions (consistency) the estimator λ^* yields “first order” asymptotic optimality, see Theorem 1 in Glynn and Szechtman (2002).

Another very significant advance in this area, is the relaxation of the equality constraint on the control variate, namely, $E[Y] = \nu$, examined in Szechtman and Glynn (2001). Notably, they provide a nonparametric maximum likelihood methodology such that an analyst may leverage inequality constraints, $E[Y] \leq \nu$ for example, to reduce the variance of their estimate for μ similar to the standard control variates method. Further, they find that the large-sample bevahior of their proposed nonparametric maximum likelihood methodology, in the case of equality constraints ($E[Y] = \nu$), is asymptotically related to control variates. This work is also very closely related to this dissertation, but again, it is in an uncensored setting.

2.5 Stratification

One final thread of research that ties into this dissertation is the concept of stratification. Two key uses of stratification that should be mentioned here are the work of Fygenson et al. (1994) and Zheng and Glynn (2017). The first is discussed in Section 2.3 but their use of stratification is with creating strata (local regions) from which the right-censoring times are drawn. Namely, each strata has its own distribution for that set of censoring times. The second use of stratification that is more closely related to this dissertation, is the work by Zheng and Glynn (2017). They create an infinite stratification of the support for X when computing $\mu = E[X]$. Summarized from Zheng and Glynn (2017), consider partitioning the underlying sample space, Ω, of X into events $A_1, A_2, \ldots$ with known $p_i \triangleq P(A_i)$ for $i \geq 1$ (note that they assume the p_i's are known, whereas this dissertation assumes they are unknown). Let the indicator function random variable $I_i = I(A_i)$ denoting the event A_i and let $(X_1, I_{i1} : i \geq 1), (X_2, I_{i2} : i \geq 1), \ldots$ be an iid sequence of samples of $(X, I_i : i \geq 1)$. This enables the following definitions and visual connection between the pre-stratification estimator $\bar{X}_n$ and the post-stratification estimator $\mathcal{P}_n$. Letting $N_n(i) \triangleq \sum_{j=1}^{n} I_{ij}$, the count of event A_i occurring in the first n samples, the pre-stratification estimator for μ is,

$$\bar{X}_n \triangleq \frac{1}{n}\sum_{j=1}^{n} X_j = \sum_{i=1}^{\infty} I\big(N_n(i) \geq 1\big) \cdot \frac{N_n(i)}{n} \cdot \frac{\sum_{j=1}^{n} X_j I_{ij}}{N_n(i)}.$$

Since the idea of stratification is to leverage knowledge of the strata, specifically the values of the p_i's, the post-stratification estimator replaces $\frac{N_n(i)}{n}$ with p_i resulting in,

$$\mathcal{P}_n \triangleq \sum_{i=1}^{\infty} p_i \cdot \frac{\sum_{j=1}^{n} X_j I_{ij}}{N_n(i)} \cdot I\big(N_n(i) \geq 1\big).$$

Of note, Asmussen and Glynn (2007) demonstrate that the post-stratification estimator, $\mathcal{P}_n$, with a finite number of strata, reduces the variance of the estimate. Zheng and Glynn (2017) extends this to the case of infinitely many strata as briefly looked at here and further provide a CLT for this reduced variance estimator. While these notable results serve as motivation for this dissertation, the key difference is that this dissertation looks at the case where the p_i's for the strata are completely unknown and where there is censoring.

Related Work

A loosely related work that served as the initial motivation for this effort is the problem examined by Akin (2017). This problem is represented by the nodes in a network that a probabilistic entity (target) can travel between. Specifically, let there be m nodes or states that the target can travel between. At each time period, the target will either stay or travel from its current node, state, or cell to some other cell connected to its current cell. The searcher, on the other hand, must select (or is randomly given, similar to this dissertation) a cell or set of cells in which to "look" or place a watcher (sensor). If the searcher knows the current location of the target and we assume that the target's behavior can be modeled as a Markov Chain, then the problem of estimating a single row of transition probabilities for this Markov Chain is a subset or special case of the problem examined in this dissertation. If the target is in say cell k then the transition probabilities $p_{k,1}, \ldots, p_{k,i}, \ldots, p_{k,m}$ form a multinomial probability distribution. Further, the target essentially samples from Z_k (the specific distribution governing the target's behavior in departing from cell k) and that sample is denoted as $Z_{k,t}$ for time period t given the searcher observes the right cell. The cell or cells in which the searcher looks are $B_{k,t}$.

This setting is both a generalization and a specific case of the primary one examined in this dissertation. Specifically, this dissertation looks at a single random variable Z, while Akin (2017) looks at multiple discrete Z's. Of note, all of the estimators examined in this dissertation form estimates for the p_i's as part of the estimation of μ, the mean of Z. Therefore, we have the same basic censoring setting as described in Section 1.1 except that the "in-strata" variance σ_i^2 is zero since each Z_k is discrete. For a much more in depth examination of this problem setting see Akin (2017).

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Theoretical Results

This chapter examines the theoretical results for the problem of point estimation in a censored environment, as described in Chapter 1. First we introduce the estimation problem with censored observations, along with three different estimating procedures. The rest of the chapter is devoted to analyzing the distribution of the resulting estimators, and to establishing connections between them. The main result from a methodological perspective is that the proposed procedures vastly improve the estimation efficiency with little extra computational overhead.

3.1 General Framework

Consider a random variable Z with unknown mean $\mu = E[Z]$ and underlying sample space Ω. The goal is to estimate μ via Monte Carlo simulation, where the samples of Z are censored. The censoring occurs as follows. There exists a random variable B independent of Z that takes values over a finite partition $\{c_1, c_2, \ldots, c_m\}$ of Ω, with $m < \infty$. In this way, Ω is split into m strata and B returns a subset of $\{c_1, c_2, \ldots, c_m\}$. Pairs (Z_t, B_t) are collected in iid fashion for each realization $t = 1, \ldots, T$, and the value of Z_t is observed if B_t contains the stratum where Z_t lies; otherwise, one can only conclude that Z_t takes a value not in the strata contained in B_t. The key motivating question is: How can the information gained by the censored samples be used to improve the estimator efficiency?

The crude estimator of μ is based off the computation of the sample mean of $Z_1, \ldots, Z_T$ over each strata for the uncensored samples, and the estimates of the strata probabilities. More specifically, the within-stratum sample mean by realization T is

$$\bar{Z}_{iT} = \frac{\sum_{t=1}^{T} Z_t I(Z_t \in c_i, c_i \in B_t)}{N_{iT}} I(N_{iT} \geq 1),$$

and

$$\bar{p}_{iT} = \frac{N_{iT}}{\sum_{t=1}^{T} I(c_i \in B_t)} I(M_{iT} \geq 1), \tag{3.1}$$

is the estimator of p_i, where $p_i = P(Z \in c_i)$, $N_{iT} = \sum_{t=1}^{T} I(Z_t \in c_i, c_i \in B_t)$, and $M_{iT} = \sum_{t=1}^{T} I(c_i \in B_t)$.

Censored Naive (CN) Estimator
Therefore, the CN estimator of μ is

$$\bar{Z}_T^{cn} = \sum_{i=1}^{m} \bar{Z}_{iT} \bar{p}_{iT}.$$

The CN estimator essentially throws away the pairs (Z_t, B_t) when B_t does not contain the stratum where Z_t lies, and will serve as a baseline for comparison of the estimators we propose.

The censoring structure, together with the constraint that $\sum_{i \in [m]} p_i = 1$, can be used to reduce the variance of the p_i estimators, which induces variance reduction of the estimator for μ. In particular, we consider two estimators, one rooted in empirical likelihood ideas, and another motivated by the method of control variates. We introduce these two estimators in turn.

Maximum Likelihood (ML) Estimator
Consider the empirical likelihood

$$\prod_{i=1}^{m} w_{iT}^{N_{iT}} (1 - w_{iT})^{M_{iT} - N_{iT}},$$

where $\boldsymbol{w}_T = (w_{1T}, \ldots, w_{mT})$ lies in the m-dimensional simplex, $\mathcal{S}_m$. The weights $\boldsymbol{w}_T^*$ that maximize the empirical likelihood are the unique solution of the optimization problem

$$\max_{\boldsymbol{w}_T \in \mathcal{S}_m} \sum_{i=1}^{m} N_{iT} \log(w_{iT}) + (M_{iT} - N_{iT}) \log(1 - w_{iT}),$$

and the resulting estimator for μ is

$$\bar{Z}_T^{ml} = \sum_{i=1}^{m} \bar{Z}_{iT} w_{iT}^*.$$

Control Variates (CV) Estimator

The method of control variates uses the knowledge that $E\big[\sum_{k=1}^{m} \bar{p}_{kT}\big] = 1$ to reduce the crude estimator variance. The control variates estimator of μ is

$$\bar{Z}_T^{cv} = \sum_{i=1}^{m} \bar{Z}_{iT}\bar{p}_{iT}^{cv},$$

where

$$\bar{p}_{iT}^{cv} = \bar{p}_{iT} + \left(\sum_{k=1}^{m} \bar{p}_{kT} - 1\right)\eta_{iT},$$

and η_{iT} is the control variates coefficient for stratum i. The control variates coefficient η_{iT} is chosen to minimize the estimator variance.

We conclude this section with notation to be used in the ensuing developments. Let $\mathcal{B}_t$ and $\mathcal{C}_t$ be the σ-algebras induced by $\{B_1, \ldots, B_t\}$ and $\{C_1, \ldots, C_t\}$, respectively, where C_t is the stratum that contains Z_t (that is, $C_t = c_i$ if $Z_t \in c_i$). Also define the conditional mean and variance of Z for stratum i, $\mu_i = E[Z|Z \in c_i]$ and $\sigma_i^2 = \text{Var}(Z|Z \in c_i)$, respectively, and $\sigma^2 = \text{Var}(Z)$. Last, let $q_i = P(c_i \in B)$, and $p_i = P(Z \in c_i)$, for $i = 1, \ldots, m$.

Further, we define for measurable functions f and g,

$$\begin{aligned} f(x) &= O\big(g(x)\big) \text{ as } x \to \infty \text{ implies } \exists k, x_0 \text{ s.t. } |f(x)| \leq kg(x) \quad \forall x \geq x_0 \\ f(x) &= o\big(g(x)\big) \text{ as } x \to \infty \text{ implies } \forall \varepsilon\ \exists k \text{ s.t. } |f(x)| \leq \varepsilon g(x) \quad \forall x \geq k \end{aligned}$$

and, for a set of r.v.'s X_n and a corresponding set of constants a_n (indexed by n),

$$\begin{aligned} X_n &= o_p(a_n) \text{ as } n \to \infty \text{ implies } && \lim_{n\to\infty} P\left(\left|\frac{X_n}{a_n}\right| \geq \varepsilon\right) = 0 \quad \forall \varepsilon > 0 \\ X_n &= O_p(a_n) \qquad\qquad \text{implies } \exists M, n > 0 \text{ s.t. } && P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \varepsilon \quad \forall \varepsilon > 0. \end{aligned}$$

The basic sketch of the rest of this chapter is as follows. We begin by analyzing the CN estimator, followed by the CV estimator, and then the ML estimator, concluding with the connection between the CV and ML estimators.

3.2 Censored Naive Estimator

In this section we obtain a central limit theorem (CLT) for the censored naive estimator, $\bar{Z}_T^{cn}$. The starting point is a CLT for the stratum sample mean $\bar{Z}_{iT}$,

$$\sqrt{T}\left(\bar{Z}_{iT} - \mu_i\right) I\left(N_{iT} \geq 1\right) \;\Rightarrow\; N\left(0,\, \frac{\sigma_i^2}{p_i q_i}\right). \tag{3.2}$$

where $\Rightarrow$ denotes weak convergence and $N(\mu, \sigma^2)$ denotes a normal random variable with mean μ and variance σ^2.

Lemma 3.2.1. *If $p_i q_i > 0$ and $\sigma_i^2 < \infty$ then (3.2) holds.*

Let Z_i' be a random variable with distribution $P_Z(\cdot | Z \in c_i, c_i \in B)$. Now note that $\bar{Z}_{iT} \overset{D}{=} \frac{1}{N_{iT}} \sum\limits_{t=1}^{N_{iT}} Z_{it}'$ on the event $N_{iT} \geq 1$, so the result follows once it's shown that

$$\sqrt{T}\, I(N_{iT} \geq 1) \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{N_{iT}} \Rightarrow N\left(0, \frac{\sigma_i^2}{p_i q_i}\right).$$

Proof. Write

$$\sqrt{T}\, \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{N_{iT}} I\left(N_{iT} \geq 1\right) = \sqrt{T}\, \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{N_{iT}} I\left(N_{iT} \geq 1\right) - \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{\sqrt{T}\, p_i q_i} + \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{\sqrt{T}\, p_i q_i}.$$

Since $\frac{N_{iT}}{T p_i q_i} \to 1$ almost surely (by the law of large numbers), the random index CLT shows that

$$\frac{1}{\sqrt{T}} \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{p_i q_i} \Rightarrow N\left(0, \frac{\sigma_i^2}{p_i q_i}\right).$$

To complete the proof, we need to show that

$$\sqrt{T}\, \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{N_{iT}} I\left(N_{iT} \geq 1\right) - \frac{\sum\limits_{t=1}^{N_{iT}} (Z_{it}' - \mu_i)}{\sqrt{T}\, p_i q_i} \Rightarrow 0. \tag{3.3}$$

Write

$$E\left[\left(\sqrt{T}\,\frac{\sum_{t=1}^{N_{iT}}(Z'_{it}-\mu_i)}{N_{iT}}I\left(N_{iT}\geq 1\right)-\frac{\sum_{t=1}^{N_{iT}}(Z'_{it}-\mu_i)}{\sqrt{T}\,p_iq_i}\right)^2\middle|\mathcal{B}_T,\mathcal{C}_T\right]$$

$$\begin{aligned}
&=\left(\frac{\sqrt{T}}{N_{iT}}I\left(N_{iT}\geq 1\right)-\frac{1}{\sqrt{T}\,p_iq_i}\right)^2 E\left[\left(\sum_{t=1}^{N_{iT}}(Z'_{it}-\mu_i)\right)^2\middle|\mathcal{B}_T,\mathcal{C}_T\right]\\
&=\left(\frac{\sqrt{T}}{N_{iT}}I\left(N_{iT}\geq 1\right)-\frac{1}{\sqrt{T}\,p_iq_i}\right)^2 N_{iT}\sigma_i^2 \qquad \text{(definition of } \sigma_i^2\text{)}\\
&=\left(\frac{T}{N_{iT}}I\left(N_{iT}\geq 1\right)-\frac{2I\left(N_{iT}\geq 1\right)}{p_iq_i}+\frac{N_{iT}}{Tp_i^2q_i^2}\right)\sigma_i^2 \qquad (3.4)\\
&=\left(\frac{T}{N_{iT}}I\left(\frac{T}{N_{iT}}\leq\frac{1+\varepsilon}{p_iq_i}\right)-\frac{2I\left(N_{iT}\geq 1\right)}{p_iq_i}+\frac{N_{iT}}{Tp_i^2q_i^2}\right)\sigma_i^2\\
&\qquad\qquad+\frac{T}{N_{iT}}I\left(\frac{T}{N_{iT}}>\frac{1+\varepsilon}{p_iq_i};N_{iT}\geq 1\right)\sigma_i^2\\
&\leq\left(\frac{1+\varepsilon}{p_iq_i}I\left(\frac{T}{N_{iT}}\leq\frac{1+\varepsilon}{p_iq_i}\right)-\frac{2I\left(N_{iT}\geq 1\right)}{p_iq_i}+\frac{N_{iT}}{Tp_i^2q_i^2}\right)\sigma_i^2\\
&\qquad\qquad+TI\left(\frac{T}{N_{iT}}>\frac{1+\varepsilon}{p_iq_i};N_{iT}\geq 1\right)\sigma_i^2,
\end{aligned}$$

for $\varepsilon>0$.

It follows that

$$
\begin{aligned}
&E\left[\left(\sqrt{T}\,\frac{\sum\limits_{t=1}^{N_{iT}}(Z'_{it}-\mu_i)}{N_{iT}}I\left(N_{iT}\geq 1\right)-\frac{\sum\limits_{t=1}^{N_{iT}}(Z'_{it}-\mu_i)}{\sqrt{T}\,p_iq_i}\right)^2\right]\\
&\leq\left(\frac{1+\varepsilon}{p_iq_i}P\left(\frac{T}{N_{iT}}\leq\frac{1+\varepsilon}{p_iq_i}\right)-\frac{2P\left(N_{iT}\geq 1\right)}{p_iq_i}+\frac{1}{p_iq_i}\right)\sigma_i^2+T\sigma_i^2P\left(1\leq N_{iT}<\frac{Tp_iq_i}{1+\varepsilon}\right)\\
&=\left(\frac{1+\varepsilon}{p_iq_i}P\left(N_{iT}\geq\frac{Tp_iq_i}{1+\varepsilon}\right)+\frac{2P\left(N_{iT}=0\right)}{p_iq_i}-\frac{1}{p_iq_i}\right)\sigma_i^2+T\sigma_i^2P\left(1\leq N_{iT}<\frac{Tp_iq_i}{1+\varepsilon}\right)\\
&=\left(\frac{\varepsilon}{p_iq_i}-\frac{1+\varepsilon}{p_iq_i}P\left(N_{iT}<\frac{Tp_iq_i}{1+\varepsilon}\right)+\frac{2P\left(N_{iT}=0\right)}{p_iq_i}\right)\sigma_i^2+T\sigma_i^2P\left(1\leq N_{iT}<\frac{Tp_iq_i}{1+\varepsilon}\right)\\
&\leq\left(\frac{\varepsilon}{p_iq_i}-\left(\frac{1+\varepsilon}{p_iq_i}-T\right)\exp\left(-2T\left(\frac{p_iq_i\varepsilon}{1+\varepsilon}\right)^2\right)+\frac{2P\left(N_{iT}=0\right)}{p_iq_i}\right)\sigma_i^2\ \text{(Hoeffding's)}\\
&\to 0
\end{aligned}
$$

as, $T\to\infty$ for $\varepsilon=O(T^{-\eta})$ with $\eta\in\left(0,\frac{1}{2}\right)$. Markov's inequality now results in Equation (3.3). ■

The next result is a CLT for $\bar{p}_{iT}$; its proof mimics that of Lemma 3.2.1, so we just highlight the differences.

Lemma 3.2.2. *If* $q_i>0$ *then*

$$\sqrt{T}\left(\bar{p}_{iT}-p_i\right)\ \Rightarrow\ N\left(0,\ \frac{p_i(1-p_i)}{q_i}\right).$$

Proof. Let p'_i be a Bernoulli rv with $P(p'_i=1)=P(Z_t\in c_i)$. Since $\frac{M_{iT}}{Tq_i}\to 1$ almost surely, the random index CLT leads to

$$\frac{1}{\sqrt{T}}\frac{\sum\limits_{t=1}^{M_{iT}}(p'_{it}-p_i)}{q_i}\Rightarrow N\left(0,\frac{p_i(1-p_i)}{q_i}\right).$$

Also,

$$\left(\frac{\sqrt{T}}{M_{iT}} I\left(M_{iT} \geq 1\right) - \frac{1}{\sqrt{T}\, q_i}\right)^2 E\left[\left(\sum_{t=1}^{M_{iT}} (p'_{it} - p_i)\right)^2 \middle| \mathcal{B}_T\right]$$

$$\begin{aligned} &= \left(\frac{\sqrt{T}}{M_{iT}} I\left(M_{iT} \geq 1\right) - \frac{1}{\sqrt{T}\, q_i}\right)^2 M_{iT} p_i (1 - p_i) \\ &= \left(\frac{T}{M_{iT}} I\left(M_{iT} \geq 1\right) - \frac{2I\left(M_{iT} \geq 1\right)}{q_i} + \frac{M_{iT}}{T q_i^2}\right) p_i (1 - p_i), \end{aligned}$$

which is the analogue of (3.4). We conclude that

$$E\left[\left(\sqrt{T}\, \frac{\sum\limits_{t=1}^{M_{iT}} (p'_{it} - p_i)}{M_{iT}} I\left(M_{iT} \geq 1\right) - \frac{\sum\limits_{t=1}^{M_{iT}} (p'_{it} - p_i)}{\sqrt{T}\, q_i}\right)^2\right] \to 0$$

as $T \to \infty$. Markov's inequality now results in

$$\sqrt{T}\, \frac{\sum\limits_{t=1}^{M_{iT}} (p'_{it} - p_i)}{M_{iT}} I\left(M_{iT} \geq 1\right) - \frac{\sum\limits_{t=1}^{M_{iT}} (p'_{it} - p_i)}{\sqrt{T}\, q_i} \Rightarrow 0.$$

Since $\bar{p}_{iT} \overset{D}{=} \frac{1}{M_{iT}} \sum\limits_{t=1}^{M_{iT}} p'_{it}$, the claim follows. ■

Lemma 3.2.3. *If $p_i q_i > 0$ and $\sigma_i^2 < \infty$ then*

$$\sqrt{T} \begin{pmatrix} \bar{Z}_{iT} - \mu_i \\ \bar{p}_{iT} - p_i \end{pmatrix} I\left(N_{iT} \geq 1\right) \Rightarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_i^2}{p_i q_i} & 0 \\ 0 & \frac{p_i(1-p_i)}{q_i} \end{pmatrix}\right).$$

Proof. Note that

$$E[\bar{Z}_{iT}] = E[E[\bar{Z}_{iT} | \mathcal{B}_T, \mathcal{C}_T]] = \mu_i P(N_{iT} \geq 1) = \mu_i + o\Big(\frac{1}{T}\Big),$$

and

$$E[\bar{p}_{iT}] = E[E[\bar{p}_{iT}|\mathcal{B}_T]] = p_i P(N_{iT} \geq 1) = p_i + o\Big(\frac{1}{T}\Big). \tag{3.5}$$

Likewise,

$$\begin{aligned} E[\bar{Z}_{iT}\bar{p}_{iT}] &= E[E[\bar{Z}_{iT}\bar{p}_{iT}|\mathcal{B}_T, \mathcal{C}_T]] \\ &= \mu_i E\left[\frac{\sum_{t=1}^{T} I(Z_t \in c_i, c_i \in B_t)}{\sum_{t=1}^{T} I(c_i \in B_t)} I(N_{iT} \geq 1)\right] \\ &= \mu_i p_i P(N_{iT} \geq 1) \\ &= \mu_i p_i + o\Big(\frac{1}{T}\Big). \end{aligned} \tag{3.6}$$

It follows that $\operatorname{Cov}(\bar{Z}_{iT}, \bar{p}_{iT}) = o(\frac{1}{T})$. The Cramér-Wold device, Durrett (2019), and Lemmas 3.2.1–3.2.2 complete the proof. ■

Let $q_{ij} = P(c_i \in B_t, c_j \in B_t)$ and $M_{ijT} = \sum_{t=1}^{T} I(c_i \in B_t, c_j \in B_t)$. The next result concerns the asymptotic covariance between strata estimators.

Lemma 3.2.4. *If $p_i q_i > 0$ and $\sigma_i^2 < \infty$ then*

$$T\,Cov(\bar{Z}_{iT}\bar{p}_{iT}, \bar{Z}_{jT}\bar{p}_{jT}) \;\to\; -p_i p_j \mu_i \mu_j \frac{q_{ij}}{q_i q_j},$$

as $T \to \infty$.

Proof. Note that

$$E[\bar{p}_{iT}\bar{p}_{jT}] = p_i p_j \left(1 - E\left[\frac{M_{ijT}}{M_{iT}M_{jT}} I\big(M_{iT} \geq 1\big) I\big(M_{jT} \geq 1\big)\right]\right). \tag{3.7}$$

Next, we lower bound the term inside the expectation,

$$\begin{aligned} &\frac{q_{ij} - \varepsilon}{(q_i + \varepsilon)(q_j + \varepsilon)} I\left(\Big|\frac{M_{iT}}{T} - q_i\Big| \leq \varepsilon,\ \Big|\frac{M_{jT}}{T} - q_j\Big| \leq \varepsilon,\ \Big|\frac{M_{ijT}}{T} - q_{ij}\Big| \leq \varepsilon\right) \\ &\qquad\qquad\qquad\qquad \leq\ T\frac{M_{ijT}}{M_{iT}M_{jT}} I\big(M_{iT} \geq 1\big) I\big(M_{jT} \geq 1\big), \end{aligned}$$

for $\varepsilon > 0$. Therefore,

$$\begin{aligned}
&TE\Big[\frac{M_{ijT}}{M_{iT}M_{jT}} I\big(M_{iT} \geq 1\big) I\big(M_{jT} \geq 1\big)\Big] \\
&\qquad\qquad \geq \frac{(q_{ij}-\varepsilon)}{(q_i+\varepsilon)(q_j+\varepsilon)} \Bigg(1 - P\Big(\Big|\frac{M_{iT}}{T} - q_i\Big| > \varepsilon\Big)\Bigg) \\
&\qquad\qquad - \frac{(q_{ij}-\varepsilon)}{(q_i+\varepsilon)(q_j+\varepsilon)} \Bigg(P\Big(\Big|\frac{M_{jT}}{T} - q_j\Big| > \varepsilon\Big) - P\Big(\Big|\frac{M_{ijT}}{T} - q_{ij}\Big| > \varepsilon\Big)\Bigg).
\end{aligned}$$

Choosing $\varepsilon = O(T^{-\eta})$ for $\eta \in (0, \frac{1}{2})$ and Hoeffding's inequality leads to

$$TE\left[\frac{M_{ijT}}{M_{iT}M_{jT}} I(M_{iT} \geq 1) I(M_{jT} \geq 1)\right] \geq \frac{q_{ij}}{q_i q_j}\left(1 - o(1)\right).$$

A similar argument produces the upper bound

$$TE\left[\frac{M_{ijT}}{M_{iT}M_{jT}} I(M_{iT} \geq 1) I(M_{jT} \geq 1)\right] \leq \frac{q_{ij}}{q_i q_j}\left(1 + o(1)\right),$$

allowing us to conclude that

$$TE\left[\frac{M_{ijT}}{M_{iT}M_{jT}} I(M_{iT} \geq 1) I(M_{jT} \geq 1)\right] \;=\; \frac{q_{ij}}{q_i q_j} + o(1). \tag{3.8}$$

Therefore,

$$\begin{aligned}
E[\bar{p}_{iT}\bar{p}_{jT}] \;&=\; p_i p_j \left(1 - \frac{1}{T}\Big(\frac{q_{ij}}{q_i q_j} \;+\; o(1)\Big)\right) \\
&=\; p_i p_j \;-\; \frac{1}{T}\Big(p_i p_j \frac{q_{ij}}{q_i q_j}\Big) \;+\; o\Big(\frac{1}{T}\Big).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{Cov}\big(\bar{p}_{iT}, \bar{p}_{jT}\big) \;&=\; E[\bar{p}_{iT}\bar{p}_{jT}] - p_i p_j \;+\; o\Big(\frac{1}{T}\Big) \\
&=\; -\frac{1}{T} p_i p_j \frac{q_{ij}}{q_i q_j} \;+\; o\Big(\frac{1}{T}\Big),
\end{aligned} \tag{3.9}$$

and

$$\operatorname{Var}\left(\sum_{i=1}^{m} \bar{p}_{iT}\right) \;=\; \frac{1}{T}\left(\sum_{i=1}^{m} \frac{p_i(1-p_i)}{q_i}\right) - \frac{2}{T}\sum_{i\neq j} p_i p_j \frac{q_{ij}}{q_i q_j} \;+\; o\Big(\frac{1}{T}\Big).$$

Hence,

$$\begin{aligned}
&\operatorname{Cov}\Big(\bar{Z}_{iT}\bar{p}_{iT},\, \bar{Z}_{jT}\bar{p}_{jT}\Big) \\
&\qquad = E\Big[E\big[\bar{Z}_{iT}\bar{p}_{iT}\bar{Z}_{jT}\bar{p}_{jT} \mid \mathcal{B}_T, \mathcal{C}_T\big]\Big] - \mu_i\mu_j p_i p_j + o\Big(\frac{1}{T}\Big) \quad \text{(by (3.6))} \\
&\qquad = \mu_i\mu_j\Big(E\big[\bar{p}_{iT}\bar{p}_{jT}\big] - p_i p_j\Big) + o\Big(\frac{1}{T}\Big) \\
&\qquad = -p_i p_j \mu_i \mu_j E\left[\frac{M_{ijT}}{M_{iT}M_{jT}} I(M_{iT} \geq 1) I(M_{jT} \geq 1)\right] + o\Big(\frac{1}{T}\Big), \text{ (by (3.7))} \\
&\qquad = -\frac{1}{T} p_i p_j \mu_i \mu_j \frac{q_{ij}}{q_i q_j} + o\Big(\frac{1}{T}\Big) \quad \text{(by (3.8))}.
\end{aligned}$$

■

The CLT for $\bar{Z}_T^{cn}$ now easily follows from the preceding lemmas.

Theorem 3.2.5. *If $p_i q_i > 0$ for $i = 1, \ldots, m$ and $\sigma^2 < \infty$,*

$$\sqrt{T}\left(\bar{Z}_T^{cn} - \mu\right) \;\Rightarrow\; N\left(0,\, \sum_{i=1}^{m}\left(\frac{\sigma_i^2 p_i}{q_i} + \mu_i^2 \frac{p_i(1-p_i)}{q_i}\right) - 2\sum_{i\neq j} p_i p_j \mu_i \mu_j \frac{q_{ij}}{q_i q_j}\right).$$

Proof. We apply the multivariate delta method twice. First, from Lemma 3.2.3,

$$\begin{aligned}
\sqrt{T}\left(\bar{Z}_{iT}\bar{p}_{iT} - \mu_i p_i\right) \;&\Rightarrow N\left(0,\, \begin{pmatrix} p_i & \mu_i \end{pmatrix} \begin{pmatrix} \frac{\sigma_i^2}{p_i q_i} & 0 \\ 0 & \frac{p_i(1-p_i)}{q_i} \end{pmatrix} \begin{pmatrix} p_i \\ \mu_i \end{pmatrix}\right) \\
&= N\left(0,\, \frac{\sigma_i^2 p_i}{q_i} + \mu_i^2 \frac{p_i(1-p_i)}{q_i}\right),
\end{aligned} \tag{3.10}$$

and second, from (3.10) and Lemma 3.2.4,

$$\begin{aligned}\sqrt{T}\left(\bar{Z}_{iT}\bar{p}_{iT}+\bar{Z}_{jT}\bar{p}_{jT}-\mu_i p_i-\mu_j p_j\right)\\ \Rightarrow N\left(0,\ \begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}\frac{\sigma_i^2 p_i}{q_i}+\mu_i^2\frac{p_i(1-p_i)}{q_i} & -p_i p_j\mu_i\mu_j\frac{q_{ij}}{q_i q_j}\\ -p_i p_j\mu_i\mu_j\frac{q_{ij}}{q_i q_j} & \frac{\sigma_j^2 p_j}{q_j}+\mu_j^2\frac{p_j(1-p_j)}{q_j}\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}\right)\\ =N\left(0,\ \frac{\sigma_i^2 p_i}{q_i}+\mu_i^2\frac{p_i(1-p_i)}{q_i}-2p_i p_j\mu_i\mu_j\frac{q_{ij}}{q_i q_j}+\frac{\sigma_j^2 p_j}{q_j}+\mu_j^2\frac{p_j(1-p_j)}{q_j}\right).\end{aligned} \tag{3.11}$$

The claim now follows from (3.11) by an induction argument. ■

Remark that if $q_{ij}\to 1$ for all strata pairs i,j, we get

$$\begin{aligned}T\mathrm{Var}(\bar{Z}^{cn}) &\to \sum_{i=1}^{m}\left(p_i(\sigma_i^2+\mu_i^2)-p_i^2\mu_i^2\right)-2\sum_{i\neq j}p_i p_j\mu_i\mu_j+o(1)\\ &=\sum_{i=1}^{m}p_i(\sigma_i^2+\mu_i^2)-\mu^2+o(1)\\ &=\sigma^2+o(1),\end{aligned}$$

so we recover the uncensored estimator variance, as expected. In the case of $P(|B_t|=1)=1$, meaning that B_t is a singleton, we have

$$\mathrm{Var}(\bar{Z}^{cn})=\frac{1}{T}\sum_{i=1}^{m}\left(\frac{\sigma_i^2 p_i}{q_i}+\mu_i^2\frac{p_i(1-p_i)}{q_i}\right),$$

so the estimator variance is very sensitive to μ. In particular, $\mathrm{Var}(\bar{Z}^{cn})$ grows with the strata means squared μ_i^2.

3.3 Control Variates Estimator

Recall that the control variates estimator of μ is

$$\bar{Z}_T^{cv}=\sum_{i=1}^{m}\bar{Z}_{iT}\bar{p}_{iT}^{cv},$$

where

$$\bar{p}^{cv}_{iT} = \bar{p}_{iT} + \left(\sum_{k=1}^{m} \bar{p}_{kT} - 1 \right) \eta_{iT}, \tag{3.12}$$

and η_{iT} is the control variates coefficient for stratum i. The control variates coefficient for stratum i that minimizes the estimator variance is given by,

$$\begin{aligned} -\frac{\mathrm{Cov}(\bar{p}_{iT}, \sum_{k=1}^{m} \bar{p}_{kT})}{\mathrm{Var}(\sum_{k=1}^{m} \bar{p}_{kT})} &= -\frac{\mathrm{Var}(\bar{p}_{iT}) + \sum_{i \neq j} \mathrm{Cov}(\bar{p}_{iT}, \bar{p}_{jT})}{\sum_{j=1}^{m} \mathrm{Var}(\bar{p}_{jT}) + 2 \sum_{j \neq k} \mathrm{Cov}(\bar{p}_{jT}, \bar{p}_{kT})} \\ &= \underbrace{-\frac{\frac{p_i(1-p_i)}{q_i} - \sum_{i \neq j} p_i p_j \frac{q_{ij}}{q_i q_j}}{\sum_{j=1}^{m} \frac{p_j(1-p_j)}{q_j} - 2 \sum_{j \neq k} p_j p_k \frac{q_{jk}}{q_j q_k}}}_{=\eta_i} + o(1), \end{aligned} \tag{3.13}$$

by (3.5), (3.7), and (3.9). However, η_i is not implementable. Letting $\bar{q}_{iT} = \frac{1}{T} M_{iT}$ and $\bar{q}_{ijT} = \frac{1}{T} M_{ijT}$ for all $i, j \in [m]$, in practice we estimate the optimal coefficient via

$$\eta_{iT} = \begin{cases} -\frac{\frac{\bar{p}_{iT}(1-\bar{p}_{iT})}{\bar{q}_{iT}} - \sum_{i \neq j} \bar{p}_{iT} \bar{p}_{jT} \frac{\bar{q}_{ijT}}{\bar{q}_{iT} \bar{q}_{jT}}}{\sum_{j=1}^{m} \frac{\bar{p}_{jT}(1-\bar{p}_{jT})}{\bar{q}_{jT}} - 2 \sum_{j \neq k} \bar{p}_{jT} \bar{p}_{kT} \frac{\bar{q}_{jkT}}{\bar{q}_{jT} \bar{q}_{kT}}}, & \text{if } \bar{p}_{\ell T} > 0, \ \forall \ell \in [m] \\ 0, & \text{otherwise.} \end{cases}$$

By the law of large numbers and continuous mapping we have $\eta_{iT} = n_i + o(1)$ a.s. Since the variance reduction induced by control variates is one minus the squared correlation coefficient between $\bar{p}_{iT}$ and $\sum_{k=1}^{m} \bar{p}_{kT}$, we get

$$\begin{aligned} \mathrm{Var}(\bar{p}^{cv}_{iT}) &= \mathrm{Var}(\bar{p}_{iT}) \left(1 - \frac{\mathrm{Cov}^2(\bar{p}_{iT}, \sum_{k=1}^{m} \bar{p}_{kT})}{\mathrm{Var}(\bar{p}_{iT}) \mathrm{Var}(\sum_{k=1}^{m} \bar{p}_{kT})} \right) + o(1/T) \\ &= \mathrm{Var}(\bar{p}_{iT}) - \eta_i^2 \mathrm{Var} \left(\sum_{k=1}^{m} \bar{p}_{kT} \right) + o(1/T) \\ &= \frac{1}{T} \frac{p_i(1-p_i)}{q_i} - \frac{\eta_i^2}{T} \left(\sum_{j=1}^{m} \frac{p_j(1-p_j)}{q_j} - 2 \sum_{j \neq k} p_j p_k \frac{q_{jk}}{q_j q_k} \right) + o(1/T). \end{aligned}$$

Furthermore

$$\begin{aligned}
E\big[\bar{p}_{iT}^{cv}\bar{p}_{jT}^{cv}\big] &= E\big[\bar{p}_{iT}\bar{p}_{jT}\big] + \eta_i\eta_j \mathrm{Var}\left(\sum_{k=1}^{m}\bar{p}_{kT}\right) \\
&\qquad + \eta_j E\left[\bar{p}_{iT}\bigg(\sum_{k=1}^{m}\bar{p}_{kT} - 1\bigg)\right] + \eta_i E\left[\bar{p}_{jT}\bigg(\sum_{k=1}^{m}\bar{p}_{kT} - 1\bigg)\right] + o\Big(\frac{1}{T}\Big) \\
&= E[\bar{p}_{iT}\bar{p}_{jT}] + \eta_i\eta_j \mathrm{Var}\left(\sum_{k=1}^{m}\bar{p}_{kT}\right) \\
&\qquad + \eta_j \mathrm{Cov}\left(\bar{p}_{iT}, \sum_{k=1}^{m}\bar{p}_{kT}\right) + \eta_i \mathrm{Cov}\left(\bar{p}_{jT}, \sum_{k=1}^{m}\bar{p}_{kT}\right) + o\Big(\frac{1}{T}\Big) \\
&= E[\bar{p}_{iT}\bar{p}_{jT}] - \eta_i\eta_j \mathrm{Var}\left(\sum_{k=1}^{m}\bar{p}_{kT}\right) + o\Big(\frac{1}{T}\Big) \\
&= p_i p_j \left(1 - \frac{1}{T}\frac{q_{ij}}{q_i q_j}\right) - \eta_i\eta_j \mathrm{Var}\left(\sum_{k=1}^{m}\bar{p}_{kT}\right) + o\Big(\frac{1}{T}\Big)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&T\mathrm{Cov}\Big(\bar{Z}_{iT}\bar{p}_{iT}^{cv}, \bar{Z}_{jT}\bar{p}_{jT}^{cv}\Big) \\
&\qquad = T\mu_i\mu_j(E[\bar{p}_{iT}^{cv}\bar{p}_{jT}^{cv}] - p_i p_j) + o(1) \\
&\qquad = T\mu_i\mu_j\left(E[\bar{p}_{iT}\bar{p}_{jT}] - \eta_i\eta_j \mathrm{Var}\left(\sum_{k=1}^{m}\bar{p}_{kT}\right) - p_i p_j\right) + o(1) \\
&\qquad = -T\mu_i\mu_j\left(p_i p_j - E[\bar{p}_{iT}\bar{p}_{jT}] + \eta_i\eta_j \mathrm{Var}\left(\sum_{k=1}^{m}\bar{p}_{kT}\right)\right) + o(1) \\
&\qquad = -T\mu_i\mu_j\left(p_i p_j - p_i p_j\left(1 - \frac{1}{T}\frac{q_{ij}}{q_i q_j}\right) + \eta_i\eta_j \mathrm{Var}\left(\sum_{k=1}^{m}\bar{p}_{kT}\right)\right) + o(1) \\
&\qquad = -\mu_i\mu_j\left(p_i p_j \frac{q_{ij}}{q_i q_j} + \eta_i\eta_j\left(\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k} - 2\sum_{\ell \neq k} p_\ell p_k \frac{q_{\ell k}}{q_\ell q_k}\right)\right) + o(1).
\end{aligned}$$

Using the preceding developments, we prove this section's main result, a CLT for $\bar{Z}_T^{cv}$.

Theorem 3.3.1. *If $p_i q_i > 0$ for $i = 1, \ldots, m$ and $\sigma^2 < \infty$,*

$$\sqrt{T}\left(\bar{Z}_T^{cv} - \mu\right) \Rightarrow N\left(0,\ \sum_{i=1}^{m}\left(\frac{\sigma_i^2 p_i}{q_i} + \mu_i^2 \frac{p_i(1-p_i)}{q_i}\right) - 2\sum_{i\neq j} p_i p_j \mu_i \mu_j \frac{q_{ij}}{q_i q_j}\right.$$
$$\left. - \left(\sum_{i=1}^{m} \mu_i \eta_i\right)^2 \left(\sum_{k=1}^{m} \frac{p_k(1-p_k)}{q_k} - 2\sum_{\ell\neq k} p_\ell p_k \frac{q_{\ell k}}{q_\ell q_k}\right)\right).$$

Proof. The analogue of (3.10) is

$$\sqrt{T}\left(\bar{Z}_{iT}\bar{p}_{iT}^{cv} - \mu_i p_i\right)$$
$$\Rightarrow N\left(0,\ \frac{\sigma_i^2 p_i}{q_i} + \mu_i^2\left(\frac{p_i(1-p_i)}{q_i} - \eta_i^2\left(\sum_{j=1}^{m}\frac{p_j(1-p_j)}{q_j} - 2\sum_{j\neq k} p_j p_k \frac{q_{jk}}{q_j q_k}\right)\right)\right).$$

Then, as in (3.11), we obtain that asymptotically $\sqrt{T}\left(\bar{Z}_{iT}\bar{p}_{iT}^{cv} + \bar{Z}_{jT}\bar{p}_{jT}^{cv} - \mu_i p_i - \mu_j p_j\right)$ is normally distributed with mean zero and variance

$$\frac{\sigma_i^2 p_i}{q_i} + \frac{\sigma_j^2 p_j}{q_j} + \mu_i^2 \frac{p_i(1-p_i)}{q_i} + \mu_j^2 \frac{p_j(1-p_j)}{q_j}$$
$$- 2\mu_i\mu_j\left(p_i p_j \frac{q_{ij}}{q_i q_j} + \eta_i \eta_j\left(\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k} - 2\sum_{\ell\neq k} p_\ell p_k \frac{q_{\ell k}}{q_\ell q_k}\right)\right)$$
$$- \left(\eta_i^2\mu_i^2 + \eta_j^2\mu_j^2\right)\left(\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k} - 2\sum_{\ell\neq k} p_\ell p_k \frac{q_{\ell k}}{q_\ell q_k}\right)$$
$$= \frac{\sigma_i^2 p_i}{q_i} + \frac{\sigma_j^2 p_j}{q_j} + \mu_i^2 \frac{p_i(1-p_i)}{q_i} + \mu_j^2 \frac{p_j(1-p_j)}{q_j} - 2\mu_i\mu_j p_i p_j \frac{q_{ij}}{q_i q_j}$$
$$- \left(\eta_i\mu_i + \eta_j\mu_j\right)^2\left(\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k} - 2\sum_{\ell\neq k} p_\ell p_k \frac{q_{\ell k}}{q_\ell q_k}\right).$$

The result now follows by an induction argument. ■

The take away from Theorem 3.3.1 is that the variance reduction induced by control variates in relation to the censored naive estimator is

$$\left(\sum_{i=1}^{m}\mu_i\eta_i\right)^2\left(\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k}-2\sum_{\ell\neq k}p_\ell p_k\frac{q_{\ell k}}{q_\ell q_k}\right),$$

as $T\to\infty$, so that control variates can only reduce the estimator variance.

Recall that the censored naive estimator has poor performance when $B_1,\ldots,B_T$ are singletons. In the CV case with B a singleton we get

$$\begin{aligned}T\mathrm{Var}\bar{Z}_T^{cv}&=\sum_{i=1}^{m}\left(\frac{\sigma_i^2p_i}{q_i}+\mu_i^2\frac{p_i(1-p_i)}{q_i}\right)-\left(\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k}\right)\left(\sum_{i=1}^{m}\mu_i\eta_i\right)^2+o(1)\\&=\sum_{i=1}^{m}\left(\frac{\sigma_i^2p_i}{q_i}+\mu_i^2\frac{p_i(1-p_i)}{q_i}\right)-\frac{\left(\sum_{i=1}^{m}\mu_i\frac{p_i(1-p_i)}{q_i}\right)^2}{\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k}}+o(1),\end{aligned}$$

after replacing for the values of η_i and η_j, since $q_{ij}=0$ for all pairs i,j. Expanding the square and canceling common terms results in

$$T\mathrm{Var}\bar{Z}_T^{cv}=\sum_{i=1}^{m}\frac{\sigma_i^2p_i}{q_i}+\sum_{i\neq j}(\mu_i-\mu_j)^2\frac{\frac{p_i(1-p_i)}{q_i}\frac{p_j(1-p_j)}{q_j}}{\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k}}+o(1),\tag{3.14}$$

which shows that the CV estimator variance is robust to large values of μ when the strata means are close to each other (i.e., $(\mu_i-\mu_j)^2$ are small).

3.4 Maximum Likelihood Estimator

Recall that the ML estimator is the solution of

$$\max_{\boldsymbol{w}_T\in\mathcal{S}_m}\sum_{i=1}^{m}N_{iT}\log(w_{iT})+(M_{iT}-N_{iT})\log(1-w_{iT}).\tag{3.15}$$

Each summand of the objective function in (3.15) is concave, so that the overall objective function is concave. Furthermore, the constraint set $\mathcal{S}_m$ is convex. It follows that the "first-order" conditions are sufficient for optimality. In particular, when $\sum_{i=1}^{m}\bar{p}_{iT}=1$, we get that $\bar{p}_{iT}$ is a critical point (cf., (3.1)), so that $w_{iT}^*=\bar{p}_{iT}$ is the unique optimal solution of problem (3.15). The more interesting case where the random element $\sum_{i=1}^{m} \bar{p}_{iT}$ does not equal 1 is addressed in the next result.

Proposition 3.4.1. *On the event* $\{\sum_{i=1}^{m} \bar{p}_{iT} \neq 1, M_{iT} > N_{iT}, \forall i \in [m]\}$*, the solution to problem (3.15) is an equation for* w_{iT} *in terms of a Lagrange multiplier,* λ_T*,*

$$w_{iT} \;=\; \frac{M_{iT} + \lambda_T - \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}}{2\lambda_T} \tag{3.16}$$

with λ_T *being the unique solution of a root equation (that exists),*

$$g(\lambda_T) \;\triangleq\; (m-2)\lambda_T + M_T - \sum_{i=1}^{m} \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}\,, \tag{3.17}$$

where $M_T = \sum_{i=1}^{m} M_{iT}$*.*

Proof. The first-order conditions lead to

$$\lambda_T \;=\; \frac{N_{iT}}{w_{iT}} - \frac{M_{iT} - N_{iT}}{1 - w_{iT}}, \qquad\qquad \forall i \in [m] \tag{3.18}$$

which, for all $i = 1, \ldots, m$, can be rearranged and simplified in the following way,

$$\lambda_T w_{iT}(1 - w_{iT}) \;=\; N_{iT} - N_{iT} w_{iT} - w_{iT}(M_{iT} - N_{iT})$$

and therefore,

$$\lambda_T w_{iT}(1 - w_{iT}) \;=\; N_{iT} - M_{iT} w_{iT}. \tag{3.19}$$

Since $\sum_{i=1}^{m} \bar{p}_{iT} \neq 1$ the constraint on $\boldsymbol{w}$ is active and the Lagrange Multiplier, λ_T, is also active. Specifically, λ_T is non-zero and is acting to inflate and deflate the w_{iT}'s such that the constraint on $\boldsymbol{w}_T$ is met. To solve the MLE problem, rewrite Equation (3.19) as,

$$\lambda_T (w_{iT})^2 - (M_{iT} + \lambda_T) w_{iT} + N_{iT} = 0,$$

resulting in the following quadratic root equation with two roots for each w_{iT}, given by:

$$w_{iT} \;=\; \frac{M_{iT} + \lambda_T \pm \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}}{2\lambda_T}. \tag{3.20}$$

To determine which root, first examine the positive root when the Lagrange multiplier is either positive or negative. Since $M_{iT} \geq N_{iT}$, if $\lambda_T > 0$ then,

$$(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT} \;\; \geq \;\; (N_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT} \;\; = \;\; (N_{iT} - \lambda_T)^2 \;\; \geq \;\; 0$$

which implies that the positive root of Equation (3.20),

$$w_{iT} \;=\; \frac{M_{iT} + \lambda_T + \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}}{2\lambda_T} \;>\; \frac{1}{2}$$

meaning that the $\sum_i w_{iT} > 1$, breaking the $\sum_i w_{iT} = 1$ constraint.

But, if $\lambda_T < 0$, and since one can write $\lambda_T = -|\lambda_T|$ and $(-M_{iT} + |\lambda_T|)^2 = (M_{iT} - |\lambda_T|)^2$, then

$$(-M_{iT} + |\lambda_T|)^2 + 4|\lambda_T|N_{iT} \;\; \geq \;\; (-N_{iT} + |\lambda_T|)^2 + 4|\lambda_T|N_{iT} \;\; = \;\; (N_{iT} + |\lambda_T|)^2$$

which implies that the positive root of Equation (3.20),

$$\begin{aligned} w_{iT} &= \frac{-M_{iT} + |\lambda_T| - \sqrt{(-M_{iT} + |\lambda_T|)^2 + 4|\lambda_T|N_{iT}}}{2|\lambda_T|} \\ &\leq \frac{-M_{iT} + |\lambda_T| - \sqrt{(-N_{iT} + |\lambda_T|)^2 + 4|\lambda_T|N_{iT}}}{2|\lambda_T|} \\ &= \frac{-M_{iT} + |\lambda_T| - \sqrt{(N_{iT} + |\lambda_T|)^2}}{2|\lambda_T|} \\ &< 0. \end{aligned}$$

breaking the variable constraint, namely that $w_{iT} \geq 0, \forall i = 1, \ldots, m$. Hence, the negative root of Equation (3.20) is correct and therefore,

$$w_{iT} \;=\; \frac{M_{iT} + \lambda_T - \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}}{2\lambda_T}.$$

Summing Equation (3.16) across i, and using the $\sum\limits_{i=1}^{m} w_{iT} = 1$ constraint, leads to a root equation, $g(\lambda_T) = 0$, specifically,

$$\begin{aligned}
2\lambda_T \sum_{i=1}^{m} w_{iT} &= \sum_{i=1}^{m} \Big(M_{iT} + \lambda_T - \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}} \Big) \\
2\lambda_T &= m\lambda_T + \sum_{i=1}^{m} \Big(M_{iT} - \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}} \Big) \\
0 &= (m-2)\lambda_T + M_T - \sum_{i=1}^{m} \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}\,.
\end{aligned}$$

The root equation is therefore,

$$g(\lambda_T) \;=\; (m-2)\lambda_T + M_T - \sum_{i=1}^{m} \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}\,.$$

The term inside the radical, $(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}$, is always positive since, by the assumptions, if $\lambda_T > 0$ then,

$$(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT} \;>\; (M_{iT}^2 + \lambda_T^2 + 2M_{iT}\lambda_T) - 4\lambda_T M_{iT} \;=\; (M_{iT} - \lambda_T)^2 \;\geq\; 0$$

and, if $\lambda_T < 0$ then,

$$(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT} \;=\; (M_{iT} + \lambda_T)^2 \;+\; 4|\lambda_T| N_{iT} \;>\; 0.$$

To show that $g(\lambda_T)$ has a unique, non-zero root, first note that it is concave since its second derivative is strictly negative,

$$g''(\lambda_T) \;=\; -4 \sum_{i=1}^{m} \frac{N_{iT}(M_{iT} - N_{iT})}{\Big((M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT} \Big)^{\frac{3}{2}}} \;<\; 0 \qquad \forall \lambda_T \in \mathbb{R}.$$

Second, notice that,

$$g'(0) \;=\; (m-2) - \sum_{i=1}^{m} \frac{(M_{iT} - 2N_{iT})}{M_{iT}} \;=\; -2 + 2\sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} \;=\; 2\left(-1 + \sum_{i=1}^{m} \bar{p}_{iT} \right)$$

which is positive for $\sum\limits_{i=1}^{m} \bar{p}_{iT} > 1$ (corresponding to $\lambda_T > 0$), negative for $\sum\limits_{i=1}^{m} \bar{p}_{iT} < 1$ (corresponding to $\lambda_T < 0$), and zero for $\sum\limits_{i=1}^{m} \bar{p}_{iT} = 1$ (which does not happen, by assumption). Hence, the root does exist and due to the concavity of $g(\lambda_T)$, it is a unique root concluding the proof for this proposition. ■

While solving a root equation is not particularly difficult using the Bisection method for example, its computational complexity or number of iterations is of order $O\big(\log(\epsilon_0) - \log(\epsilon)\big)$ where, ϵ_0 is the initial interval size and ϵ is the tolerance or desired precision. As Equation (3.25) and Proposition 3.4.3 show, this initial interval could be as wide as $\big(0, \sqrt{T}\,\big)$ and therefore, for a fixed precision, the overall Bisection method computational cost is roughly of order $O\big(\log\big(\sqrt{T}\,\big)\big)$. We therefore seek an approximation for the Lagrange multiplier that will remove the need for root solving. Our goal is to reduce the problem's computational complexity from something along the lines of Newton's method or the Bisection method to roughly km operations (k being a constant of size less than 50 and m as defined by this problem). Hence, a constant computational cost is clearly desirable versus one that grows as the logarithm of the square root of the number of samples collected.

From Equation (3.17) we can find an approximate root by taking a Taylor Expansion of the root equation around zero. This results in the following,

$$0 \;=\; g(\lambda_T) \;=\; g(0) \;+\; \lambda_T g'(0) \;+\; \frac{\lambda_T^2}{2!} g''(0) \;+\; \frac{\lambda_T^3}{3!} g'''(\eta) \qquad \text{for } \; |\eta| \leq |\lambda_T|.$$

The first two terms of this expansion, evaluated at zero, are:

$$g'(0) \;=\; -2 + 2\sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} \qquad\qquad g''(0) \;=\; -4\sum_{i=1}^{m} \left[\frac{N_{iT}}{M_{iT}^2} - \frac{N_{iT}^2}{M_{iT}^3}\right].$$

This further leads to the following exact and approximate roots, since $g(0) = 0$ (from Equation (3.17) by inspection),

$$\lambda_T = \tilde{\lambda}_T \;-\; \frac{\lambda_T^2 g'''(\eta)}{3g''(0)}$$

where

$$\tilde{\lambda}_T = \begin{cases} -2\frac{g'(0)}{g''(0)} = \frac{M_T \sum_{i=1}^{m} \left[\frac{N_{iT}}{M_{iT}} - p_i\right]}{\sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} \left[1 - \frac{N_{iT}}{M_{iT}}\right] \frac{M_T}{M_{iT}}} & \text{if } -1 + \sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} \neq 0 \\ 0 & \text{o/w.} \end{cases} \tag{3.21}$$

Using both the Law of the Iterated Logarithm and the Strong Law of Large Numbers allows us to bound the growth of this approximate root.

Proposition 3.4.2. *Suppose* $p_i q_i > 0$ *for all* $i \in [m]$. *Then*

$$T^{-\delta} \left|\lambda_T - \tilde{\lambda}_T\right| = O(1), \qquad \textit{a.s., for } \delta \in \left(0, \frac{1}{2}\right).$$

Proof. The case $\sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} = 1$ is immediate since λ_T, being the Lagrange multiplier of Problem (3.15), equals zero when $\sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} = 1$. Therefore, this proof is broken into two primary steps when $\sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} \neq 1$. First, we establish the asymptotic growth rate of $\tilde{\lambda}_T$. Second, we show that the root equation changes signs at the endpoints of an interval centered around $\tilde{\lambda}_T$ and that that interval is of smaller order than $\tilde{\lambda}_T$. These two steps allow us to conclude the proof.

By the strong law of large numbers and continuous mapping theorem (Durrett 2019),

$$\sum_{i=1}^{m} \frac{N_{iT}}{M_{iT}} \left[1 - \frac{N_{iT}}{M_{iT}}\right] \frac{M_T}{M_{iT}} \to \sum_{i=1}^{m} p_i(1 - p_i)\frac{1}{q_i}, \tag{3.22}$$

a.s. as $T \to \infty$. Examining the numerator of Equation (3.21), the law of the iterated logarithm leads to

$$\begin{aligned} M_T - T\sum_{i=1}^{m} q_i &= O\left(\sqrt{T \log\log T}\right) \\ \bar{p}_{iT} - p_i &= O\left(\sqrt{\frac{\log\log T}{T}}\right) \end{aligned} \tag{3.23}$$

and therefore,

$$M_T \sum_{i=1}^{m} (\bar{p}_{iT} - p_i) \ = \ O\Big(\sqrt{T \log\log T}\Big) \qquad \text{a.s.} \tag{3.24}$$

Equations (3.22) and (3.24) therefore imply that,

$$\tilde{\lambda}_T = O\Big(\sqrt{T \log\log T}\Big) \qquad \text{a.s.} \tag{3.25}$$

Because the root equation, Equation (3.17), is continuous, to prove the existence of a root, it is sufficient to show that for a certain interval its endpoints differ in sign. Specifically, we will show that $g(\tilde{\lambda}_T - \beta_T)$ and $g(\tilde{\lambda}_T + \beta_T)$ are of opposite sign, for T large and β_T of smaller order than $\tilde{\lambda}_T$. Therefore, it is only necessary to examine g on a sufficiently small interval around $\tilde{\lambda}_T$. Specifically, we focus on $g(\xi)$ for $\xi \in I_T = \Big[\tilde{\lambda}_T - T^{\delta}, \tilde{\lambda}_T + T^{\delta}\Big]$ with $\delta \in \big(0, \frac{1}{2}\big)$.

We use the Taylor Theorem with Remainder around the dominating term, M_{iT}^2, in the square root term of Equation (3.17). This results in,

$$\begin{aligned} \sqrt{(M_{iT} + \xi)^2 - 4\xi N_{iT}} \ = \ M_{iT} \ &+ \ \frac{\xi(\xi + 2M_{iT} - 4N_{iT})}{2M_{iT}} \\ &- \ \frac{\xi^2(\xi + 2M_{iT} - 4N_{iT})^2}{8M_{iT}^3} \\ &+ \ \frac{\xi^3(\xi + 2M_{iT} - 4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}} \end{aligned} \tag{3.26}$$

where

$$\begin{aligned} &\gamma_T \in \big[M_{iT}^2,\ M_{iT}^2 + \xi(\xi + 2M_{iT} - 4N_{iT})\big] && \text{if} \qquad \xi(\xi + 2M_{iT} - 4N_{iT}) \ \geq \ 0 \\ &\gamma_T \in \big[M_{iT}^2 + \xi(\xi + 2M_{iT} - 4N_{iT}),\ M_{iT}^2\big] && \text{otherwise.} \end{aligned}$$

Now, plugging Equation (3.26) back into g yields that

$$\begin{aligned}
g(\xi) \;=\; & (m-2)\xi \;+\; M_T \;-\; \sum_{i=1}^{m}\Bigg[M_{iT} \;+\; \frac{\xi(\xi+2M_{iT}-4N_{iT})}{2M_{iT}} \\
& \qquad\qquad -\; \frac{\xi^2(\xi+2M_{iT}-4N_{iT})^2}{8M_{iT}^3} \;+\; \frac{\xi^3(\xi+2M_{iT}-4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}\Bigg] \\
=\; & -2\xi - \sum_{i=1}^{m}\Bigg[-\frac{2\xi N_{iT}}{M_{iT}} + \frac{4\xi^2 M_{iT}^2}{8M_{iT}^3} \\
& \quad - \frac{\xi^4 + 4\xi^2 M_{iT}^2 + 16\xi^2 N_{iT}^2 + 4\xi^3 M_{iT} - 8\xi^3 N_{iT} - 16\xi^2 M_{iT} N_{iT}}{8M_{iT}^3} \\
& \quad + \frac{\xi^3(\xi+2M_{iT}-4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}\Bigg] \\
=\; & 2\xi\left(-1 + \sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}}\right) + \sum_{i=1}^{m}\Bigg[\frac{\xi^4 + 16\xi^2 N_{iT}^2 + 4\xi^3 M_{iT} - 8\xi^3 N_{iT} - 16\xi^2 M_{iT} N_{iT}}{8M_{iT}^3} \\
& \qquad - \frac{\xi^3(\xi+2M_{iT}-4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}\Bigg].
\end{aligned}$$

Hence, the non-zero root of $g(\xi)=0$ satisfies that

$$\begin{aligned}
& 2\Bigg(-1 + \sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}}\Bigg) \\
& -\sum_{i=1}^{m}\Bigg[-\frac{\xi^3 + 16\xi N_{iT}^2 + 4\xi^2 M_{iT} - 8\xi^2 N_{iT} - 16\xi M_{iT} N_{iT}}{8M_{iT}^3} + \frac{\xi^2(\xi+2M_{iT}-4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}\Bigg]
\end{aligned}$$

is equal to zero, and hence (defining $\hat{g}(\xi)$ as this difference)

$$\begin{aligned}
\hat{g}(\xi) = & -2 + 2\sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}} + 2\xi\sum_{i=1}^{m}\left[\frac{N_{iT}(N_{iT}-M_{iT})}{M_{iT}^3}\right] \\
& + \xi^2\sum_{i=1}^{m}\left[\frac{\xi + 4M_{iT} - 8N_{iT}}{8M_{iT}^3} - \frac{(\xi+2M_{iT}-4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}\right] \\
= & \; 0.
\end{aligned}$$

With $\xi = \tilde{\lambda}_T \pm T^{\delta}$ the first three terms reduce as follows,

$$\begin{aligned}
&-2+2\sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}}+2\xi\sum_{i=1}^{m}\left[\frac{N_{iT}(N_{iT}-M_{iT})}{M_{iT}^3}\right]\\
=&-2+2\sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}}+2\tilde{\lambda}_T\sum_{i=1}^{m}\left[\frac{N_{iT}(N_{iT}-M_{iT})}{M_{iT}^3}\right]\pm 2T^{\delta}\sum_{i=1}^{m}\left[\frac{N_{iT}(N_{iT}-M_{iT})}{M_{iT}^3}\right]\\
=&-2+2\sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}}+2\left[\frac{-1+\sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}}}{\sum_{i=1}^{m}\frac{N_{iT}}{M_{iT}}\left[\frac{M_{iT}-N_{iT}}{M_{iT}^2}\right]}\right]\sum_{i=1}^{m}\left[\frac{N_{iT}(N_{iT}-M_{iT})}{M_{iT}^3}\right]\\
&\qquad\pm 2T^{\delta}\sum_{i=1}^{m}\left[\frac{N_{iT}(N_{iT}-M_{iT})}{M_{iT}^3}\right]\\
=&\pm\ 2T^{\delta}\sum_{i=1}^{m}\left[\frac{N_{iT}(N_{iT}-M_{iT})}{M_{iT}^3}\right]\\
=&\pm\ 2T^{\delta}\sum_{i=1}^{m}\frac{\bar{p}_{iT}(1-\bar{p}_{iT})}{M_{iT}}\\
=&\pm O(T^{\delta-\varepsilon_1-1}),
\end{aligned}$$

by (3.23) for $\varepsilon_1 \in (0,\delta)$.

Therefore, $\hat{g}(\tilde{\lambda}_T + T^{\delta})$ equals

$$O\big(T^{\delta-\varepsilon_1-1}\big)+(\tilde{\lambda}_T+T^{\delta})^2\sum_{i=1}^{m}\left[\frac{\tilde{\lambda}_T+T^{\delta}+4M_{iT}-8N_{iT}}{8M_{iT}^3}-\frac{(\tilde{\lambda}_T+T^{\delta}+2M_{iT}-4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}\right] \tag{3.27}$$

and $\hat{g}(\tilde{\lambda}_T - T^{\delta})$ equals

$$-O(T^{\delta-\varepsilon_1-1})+(\tilde{\lambda}_T-T^{\delta})^2\sum_{i=1}^{m}\left[\frac{\tilde{\lambda}_T-T^{\delta}+4M_{iT}-8N_{iT}}{8M_{iT}^3}-\frac{(\tilde{\lambda}_T-T^{\delta}+2M_{iT}-4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}\right] \tag{3.28}$$

To complete the proof, it remains to show that $\hat{g}(\tilde{\lambda}_T + T^{\delta}) > 0$ and $\hat{g}(\tilde{\lambda}_T - T^{\delta}) < 0$ a.s. as $T \to \infty$. To wit,

$$(\tilde{\lambda}_T \pm T^{\delta})^2\sum_{i=1}^{m}\frac{\tilde{\lambda}_T \pm T^{\delta}+4M_{iT}-8N_{iT}}{8M_{iT}^3} = O(T^{\varepsilon_2-1}) \tag{3.29}$$

for $\varepsilon_2 > 0$, by (3.23)–(3.25). Likewise, from the definition of γ_T and (3.23)–(3.25) we conclude that $\gamma_T \geq k_1 T^{2-\varepsilon_3}$ a.s. for all T sufficiently large, and for positive constants k_1, ε_3. Therefore, for $\varepsilon_4 > 0$,

$$(\tilde{\lambda}_T \pm T^\delta)^2 \sum_{i=1}^m \frac{(\tilde{\lambda}_T \pm T^\delta + 2M_{iT} - 4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}} = O(T^{\varepsilon_4 - 1}). \tag{3.30}$$

Choosing $\varepsilon_1, \varepsilon_2$ and ε_4 so that $\varepsilon_2, \varepsilon_4 < \delta - \varepsilon_1$ in (3.29)–(3.30) allows us to conclude that $\hat{g}(\tilde{\lambda}_T + T^\delta) > 0$ and $\hat{g}(\tilde{\lambda}_T - T^\delta) < 0$, meaning that the non-zero root of g lies within T^δ of $\tilde{\lambda}_T$ a.s. for all T sufficiently large on the event $\left\{ \sum\limits_{i=1}^m \frac{N_{iT}}{M_{iT}} \neq 1 \right\}$. ■

Since $\tilde{\lambda}_T = O\left(\sqrt{T \log\log T}\right)$, $\tilde{\lambda}_T$ grows faster than T^δ for $\delta \in (0, 1/2)$, meaning that the width of I_T decreases relative to $\tilde{\lambda}_T$ and therefore λ_T approaches $\tilde{\lambda}_T$. This naturally leads to the question: what is λ_T's asymptotic distribution?

The question can be answered by relating λ_T with $\tilde{\lambda}_T$. From (3.21), we can rewrite $\tilde{\lambda}_T$ as

$$\tilde{\lambda}_T = T \frac{-1 + \sum_{i=1}^m \bar{p}_{iT}}{\sum_{j=1}^m \frac{\bar{p}_{jT}(1-\bar{p}_{jT})}{q_j}}. \tag{3.31}$$

Further, the central limit theorem and (3.13) yield,

$$\sqrt{T}\,(-1 + \sum_{i=1}^m \bar{p}_{iT}) \Rightarrow N\left(0, \sum_{j=1}^m \frac{p_j(1-p_j)}{q_j} - 2\sum_{j \neq k} p_j p_k \frac{q_{jk}}{q_j q_k}\right), \tag{3.32}$$

for the numerator in (3.31). The law of large numbers and continuous mapping imply,

$$\sum_{j=1}^m \frac{\bar{p}_{jT}(1-\bar{p}_{jT})}{\bar{q}_{jT}} = \sum_{j=1}^m \frac{p_j(1-p_j)}{q_j} + o(1),$$

for the denominator in (3.31). Hence, with Slutsky's theorem, see Durrett (2019), we have that

$$\frac{1}{\sqrt{T}} \tilde{\lambda}_T \Rightarrow N\left(0, \frac{1}{\sum_{j=1}^m \frac{p_j(1-p_j)}{q_j}} - 2\frac{\sum_{j \neq k} p_j p_k \frac{q_{jk}}{q_j q_k}}{\left(\sum_{j=1}^m \frac{p_j(1-p_j)}{q_j}\right)^2}\right). \tag{3.33}$$

In light of Proposition 3.4.2, the CLT for $\tilde{\lambda}_T$ in (3.33) suggests an analogue result for λ_T. These ideas are made rigorous in the next result.

Proposition 3.4.3. *Assume that $p_i q_i > 0$ for all $i \in [m]$. Then,*

$$\frac{1}{\sqrt{T}}\lambda_T \;\Rightarrow\; N\left(0,\; \frac{1}{\sum_{j=1}^m \frac{p_j(1-p_j)}{q_j}} - 2\frac{\sum_{j\neq k} p_j p_k \frac{q_{jk}}{q_j q_k}}{\left(\sum_{j=1}^m \frac{p_j(1-p_j)}{q_j}\right)^2}\right).$$

Proof. With (3.33) in place, we only need to show that

$$\frac{1}{T}E\big[(\lambda_T - \tilde{\lambda}_T)^2\big] \;\to\; 0, \tag{3.34}$$

as $T \to \infty$, since the Markov inequality would then imply $\frac{1}{\sqrt{T}}(\lambda_T - \tilde{\lambda}_T) \Rightarrow 0$. Proposition 3.4.2 and the dominated convergence theorem result in

$$E\big[(\lambda_T - \tilde{\lambda}_T)^2\big] = O\big(T^{2\delta}\big),$$

with $\delta \in \big(0, \frac{1}{2}\big)$, whence Equation (3.34) follows. ■

Connection with Control Variates

Rewriting (3.19) as $\lambda_T w_{iT}^*(1-w_{iT}^*) = N_{iT} - M_{iT} w_{iT}^*$ leads to,

$$\begin{aligned} w_{iT}^* &= \bar{p}_{iT} - \lambda_T \frac{w_{iT}^*(1-w_{iT}^*)}{M_{iT}} \\ &\approx \bar{p}_{iT} - \tilde{\lambda}_T \frac{w_{iT}^*(1-w_{iT}^*)}{M_{iT}} \qquad \text{(Proposition 3.4.2)} \\ &= \bar{p}_{iT} + \left(1 - \sum_{i=1}^m \bar{p}_{iT}\right) \frac{\frac{w_{iT}^*(1-w_{iT}^*)}{q_i}}{\sum_{j=1}^m \frac{\bar{p}_{jT}(1-\bar{p}_{jT})}{q_j}} \qquad \text{(cf., 3.31)} \\ &\approx \bar{p}_{iT}^{cv}, \end{aligned}$$

where the last approximation (not meant to be rigorous) follows from (3.12) and (3.13) when $|B_t| = 1$. Hence, $w_{iT}^* \approx \bar{p}_{iT}^{cv}$, which motivates a connection between the ML and CV estimators.

Proposition 3.4.4. *Assume that $p_i q_i > 0$ for all $i \in [m]$. Then,*

$$w^*_{iT} = \bar{p}_{iT} + \left(1 - \sum_{i=1}^m \bar{p}_{iT}\right) \frac{\frac{p_i(1-p_i)}{q_i}}{\sum_{j=1}^m \frac{p_j(1-p_j)}{q_j}} \big(1 + o(1)\big) \qquad a.s.$$

Proof. Clearly, $w^*_{iT} = \bar{p}^{cv}_{iT}$ when $\sum_{i=1}^m \bar{p}_{iT} = 1$. On the event $\{\sum_{i=1}^m \bar{p}_{iT} \neq 1\}$ we have $\lambda_T \neq 0$ and from Proposition 3.4.1,

$$\begin{aligned} w^*_{iT} &= \frac{M_{iT}}{2\lambda_T} + \frac{\lambda_T}{2\lambda_T} - \frac{\sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}}{2\lambda_T} \\ &= \frac{M_{iT}}{2\lambda_T} + \frac{\lambda_T}{2\lambda_T} - \frac{M_{iT} + \frac{\lambda_T(\lambda_T + 2M_{iT} - 4N_{iT})}{2M_{iT}} - \frac{\lambda_T^2(\lambda_T + 2M_{iT} - 4N_{iT})^2}{8M_{iT}^3} + \frac{\lambda_T^3(\lambda_T + 2M_{iT} - 4N_{iT})^3}{16\gamma_T^{\frac{5}{2}}}}{2\lambda_T} \end{aligned}$$

from (3.26), for

$$\gamma_T \in \begin{cases} [M_{iT}^2, M_{iT}^2 + \lambda_T(\lambda_T + 2M_{iT} - 4N_{iT})] & \text{if} \quad \lambda_T(\lambda_T + 2M_{iT} - 4N_{iT}) \geq 0 \\ [M_{iT}^2 + \lambda_T(\lambda_T + 2M_{iT} - 4N_{iT}),\ M_{iT}^2] & \text{otherwise.} \end{cases}$$

Canceling terms leads, after some algebra, to

$$w^*_{iT} = \bar{p}_{iT} - \lambda_T \frac{\bar{p}_{iT}(1 - \bar{p}_{iT})}{T\bar{q}_{iT}} + \frac{\lambda_T^3}{16M_{iT}^3} + \frac{\lambda_T^2}{4M_{iT}^2} - \frac{\lambda_T^2 \bar{p}_{iT}}{2M_{iT}^2} - \frac{\lambda_T^2(\lambda_T + 2M_{iT} - 4N_{iT})^3}{32\gamma_T^{\frac{5}{2}}}.$$

Note that

$$\frac{\lambda_T^3}{16M_{iT}^3} + \frac{\lambda_T^2}{4M_{iT}^2} - \frac{\lambda_T^2 \bar{p}_{iT}}{2M_{iT}^2} - \frac{\lambda_T^2(\lambda_T + 2M_{iT} - 4N_{iT})^3}{32\gamma_T^{\frac{5}{2}}} = O(T^{\delta - 1}),$$

for $\delta \in \left(0, \frac{1}{2}\right)$, by an analysis similar to the one used in Proposition 3.4.2. Therefore,

$$w^*_{iT} = \bar{p}_{iT} - \frac{\tilde{\lambda}_T}{T} \underbrace{\frac{\bar{p}_{iT}(1 - \bar{p}_{iT})}{\bar{q}_{iT}}}_{= \frac{p_i(1-p_i)}{q_i}(1 + o(1))} + \underbrace{\frac{\tilde{\lambda}_T - \lambda_T}{T}}_{= O(T^{\delta - 1}),\text{ by Prop. 3.4.2}} \underbrace{\frac{\bar{p}_{iT}(1 - \bar{p}_{iT})}{\bar{q}_{iT}}}_{= \frac{p_i(1-p_i)}{q_i}(1 + o(1))} + O(T^{\delta - 1}).$$

The lower bound analogue of (3.25), which follows from the lim inf part of the law of iterated logarithm, is

$$\frac{\tilde{\lambda}_T}{T} \;\geq\; kT^{-1/2-\varepsilon} \qquad \text{a.s.}$$

for $\varepsilon, k > 0$ and T sufficiently large. Hence, with $\varepsilon + \delta < \frac{1}{2}$, we get

$$\begin{aligned} w^*_{iT} &= \bar{p}_{iT} - \frac{\tilde{\lambda}_T}{T}\frac{p_i(1-p_i)}{q_i}(1+o(1)) \\ &= \bar{p}_{iT} + (1 - \sum_{i=1}^{m} \bar{p}_{iT})\frac{\frac{p_i(1-p_i)}{q_i}}{\sum_{j=1}^{m}\frac{p_j(1-p_j)}{q_j}}(1+o(1)) \qquad \text{a.s.} \end{aligned}$$

by Equation (3.31). ■

We conclude this section with a central limit theorem for the ML estimator

$$\bar{Z}^{ml}_T = \sum_{i=1}^{m} \bar{Z}_{iT} w^*_{iT},$$

as defined at the beginning of this chapter.

Theorem 3.4.1. *If $p_i q_i > 0$ for $i = 1, \ldots, m$ and $\sigma^2 < \infty$,*

$$\sqrt{T}\left(\bar{Z}^{ml}_T - \mu\right) \;\Rightarrow\; N\left(0,\, \sum_{i=1}^{m}\frac{\sigma_i^2 p_i}{q_i} + \sum_{i\neq j}(\mu_i - \mu_j)^2 \frac{\frac{p_i(1-p_i)}{q_i}\frac{p_j(1-p_j)}{q_j}}{\sum_{k=1}^{m}\frac{p_k(1-p_k)}{q_k}}\right).$$

Proof. Recall from (3.14) the CV estimator variance when B_t is a singleton. Then the result is immediate from Theorem 3.3.1, Slutsky's theorem (see Durrett (2019)), and Proposition 3.4.4. ■

The main take away of this chapter is that the control variates estimator outperforms the censored naive and maximum likelihood estimators when the observations are censored. The gain is very significant when the unknown mean, $\mu = E[Z]$, is large. These results are illustrated with several numerical examples in the next chapter. Further, by using the control variates estimator one avoids needing to compute the Lagrange Multiplier from the root equation, which as mentioned earlier carries a computational cost of roughly order $O\big(\log\big(\sqrt{T}\,\big)\big)$. Even using the approximate root, $\tilde{\lambda}_T$, to inform the initial interval doesn't completely reduce this cost to a constant as that interval also grows by T^{δ} where $\delta \in \big(0, \frac{1}{2}\big)$. The control variates estimator on the other hand requires roughly km operations (k being a constant of size less than 50 and m as defined by this problem). Therefore, a constant computational cost is clearly desirable versus one that grows as the logarithm of the square root of the number of samples collected.

CHAPTER 4: Numerical Results

This chapter presents the results of numerical analysis in support of the theoretical work presented in Chapter 3. Notably, this chapter is broken into two parts, the singleton or non-combinatorial setting when $|B_t| = 1$ and the combinatorial setting when $0 < |B_t| < m$. Further, this analysis is the result of measuring the performance (estimator variance and mean squared error) of our proposed estimators compared against both the worst case estimator, censored naive, and the best case estimator, the uncensored estimator. This chapter primarily seeks to confirm via simulation the theoretical variances of the proposed estimators from Theorems 3.2.5, 3.3.1, and 3.4.1.

All simulations for this numerical analysis were conducted in the following way. A MatLab script was created that defines the parameters of the random variable Z, establishes a stratification of the associated support Ω of Z, and calculates the actual (true) population parameters (namely, the μ_i and σ_i's). Next, the script collects T iid samples from the tuple (Z, B). These samples or realizations are used to compute the values for the count variables associated with the "looks" per strata, M_{iT} and M_{ijT}, and the uncensored observations per strata, N_{iT}. These count variables combined with the uncensored Z_t's, are used to form four different estimates for the mean of Z, namely, the CN, ML, CV, and uncensored estimates (in the ML case, the estimate requires a use of MatLab's optimization software to compute the true root, λ_T). Of note, the interval given to MatLab's solver function is based on the approximate root, $\tilde{\lambda}_T$. This process is then replicated either 10^3 or 10^4 times (depending on whether it is the singleton or combinatorial section), and the sample variance and mean squared error of each estimator is then computed. These sample variances are compared against the theoretical values for a given number of samples, T, based on Theorems 3.2.5, 3.3.1, and 3.4.1.

4.1 Singleton Setting

In this section we look at the singleton setting where $|B_t| = 1$ and further, where the underlying unknown random variable being observed is $Z \sim N(\mu = 5, \sigma = 1)$.

Additionally, we define seven ($m = 7$) cell intervals using the following partitions of the real numbers: $\{4.5, 5.5, 6.5, 7.5, 8.5, 9.5\}$. This partition results in these stratum: $c_1 = (-\infty, 4.5), c_2 = (4.5, 5.5), \ldots, c_7 = (9.5, \infty)$. Further, this partition of Z, results in the specific values for the p_i's displayed in Table (4.1). Additionally, a multinomial distribution guides the external censoring algorithm with various probability masses, values for the q_i's, as displayed in Table (4.2). Of note, when the mean of Z is shifted for the simulations displayed in Table (4.7), the strata are shifted by the same amount to ensure that the p_i's remain constant.

p_1	p_2	p_3	p_4	p_5	p_6	p_7
3.085e^{-1}	3.829e^{-1}	2.417e^{-1}	6.060e^{-2}	5.977e^{-3}	2.292e^{-4}	3.398e^{-6}

Table 4.1. p_i's for shifted gaussian random variable $Z \sim N(\mu = 5, \sigma = 1)$.

Level	1 low	2 low	3 low	4 low	5 low	6 low	uniform
q_1	1.00e^{-2}	1.00e^{-2}	1.00e^{-2}	1.00e^{-2}	4.75e^{-1}	9.40e^{-1}	1.43e^{-1}
q_2	1.65e^{-1}	1.00e^{-2}	1.00e^{-2}	1.00e^{-2}	4.75e^{-1}	1.00e^{-2}	1.43e^{-1}
q_3	1.65e^{-1}	1.96e^{-1}	1.00e^{-2}	1.00e^{-2}	1.00e^{-2}	1.00e^{-2}	1.43e^{-1}
q_4	1.65e^{-1}	1.96e^{-1}	2.43e^{-1}	1.00e^{-2}	1.00e^{-2}	1.00e^{-2}	1.43e^{-1}
q_5	1.65e^{-1}	1.96e^{-1}	2.43e^{-1}	3.20e^{-1}	1.00e^{-2}	1.00e^{-2}	1.43e^{-1}
q_6	1.65e^{-1}	1.96e^{-1}	2.43e^{-1}	3.20e^{-1}	1.00e^{-2}	1.00e^{-2}	1.43e^{-1}
q_7	1.65e^{-1}	1.96e^{-1}	2.43e^{-1}	3.20e^{-1}	1.00e^{-2}	1.00e^{-2}	1.43e^{-1}

Table 4.2. Censoring algorithm's probability mass levels per strata, $m = 7$.

4.1.1 Root Equation and Distribution of Lambda

A key component of the Maximum Likelihood estimator is the calculation of the Lagrange Multiplier, λ_T, for a given set of realizations of Z. This section shows some interesting properties of that multiplier and its associated root function. As noted in Chapter 3, if $\sum\limits_{i=1}^{m} \bar{p}_{iT} = 1$, then $\lambda_T = 0$.

Root Equation or Function of Lambda

The root equation from Chapter 3, Equation (3.17), is

$$g(\lambda_T) \;=\; (m-2)\lambda_T + M_T - \sum_{i=1}^{m} \sqrt{(M_{iT} + \lambda_T)^2 - 4\lambda_T N_{iT}}$$

with a positive root if $\sum_i \bar{p}_{iT} > 1$ and a negative root if $\sum_i \bar{p}_{iT} < 1$. Note, as $\sum_i \bar{p}_{iT}$ approaches 1 from either side, the resultant root approaches zero. Figure (4.1) shows two realizations of the root function, both a positive and negative root, illustrating the function's basic shape. This simulation uses the p_i's from Table (4.1) along with the q_i's from the "2 low" column in Table (4.2), resulting in Table (4.3)'s realizations.

strata	1	2	3	4	5	6	7	$\sum_i \bar{p}_{iT}$
N_{iT}	215	283	3316	799	72	3	0	
M_{iT}	713	746	13637	13858	13540	13842	13664	0.9873
N_{iT}	212	268	3329	789	100	3	0	
M_{iT}	686	674	13691	13699	13825	13715	13710	1.0149

Table 4.3. Values for M_{iT} and N_{iT} resulting from MatLab simulations with: $T = 70k$, p_i's from Table (4.1), and q_i's from Table (4.2)'s "2 low" column.

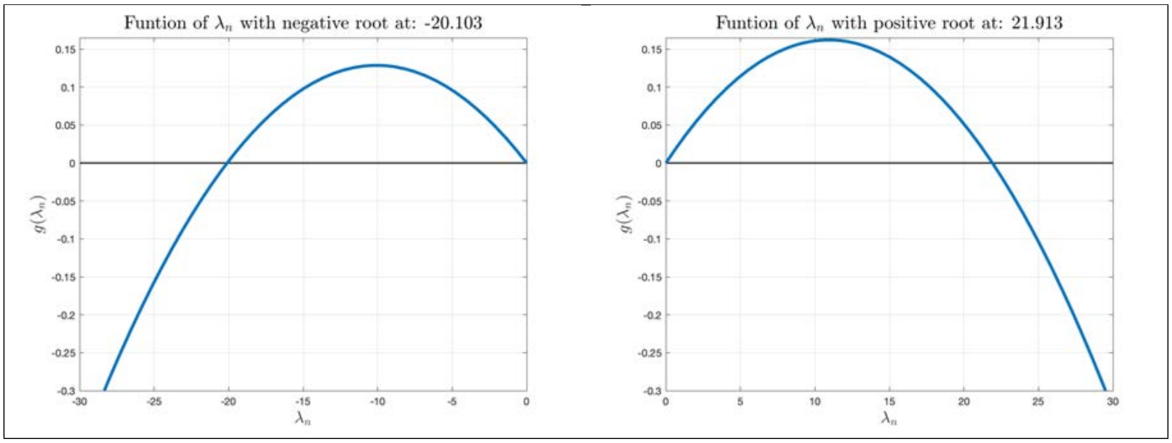

Figure 4.1. Graph of the λ_T function from MatLab for two different realizations of the simulation with specific values for M_{iT}'s and N_{iT}'s listed in Table (4.3).

Distribution of Lambda and Approximate Lambda

Here we compare the two different methods of finding the Lagrange Multiplier λ_T from the root equation from the Maximum Likelihood Estimator. The first is via the MatLab optimization software and the second is to use the Taylor Approximation to the true λ_T. Figure (4.2) displays the respective density histograms and Figure (4.3) displays the respective QQ Plots for both of these methods across 10^4 different simulations, each with $T = 70k$ with the p_i's from Table (4.1) and the q_i's from the "2 low" column in Table (4.2). As can been seen from Figures (4.2) and (4.3) both λ_T and $\tilde{\lambda}_T$ appear normal and additionally have very similar standard deviations, 38.639 and 38.752 respectively (a roughly 0.3% difference). Further, the Shapiro-Wilks test returned respective p-values of 0.149 and 0.00764 with the null hypothesis that both are normally distributed (of note, due to the large sample size, this test may be overpowered). Therefore, based on the histograms, the qq plots, and the results of the Shapiro-Wilks normality test, it appears that both both λ_T and $\tilde{\lambda}_T$ are indeed normal.

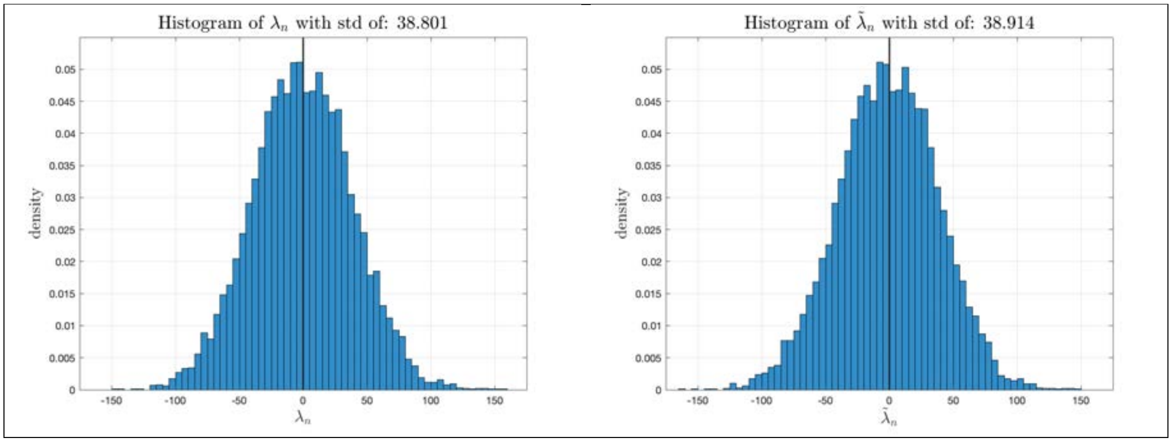

Figure 4.2. λ_T histograms for MatLab optimization software and Taylor approximation. $T = 70k$, p_i's and "2 low" q_i's from Tables (4.1) and (4.2).

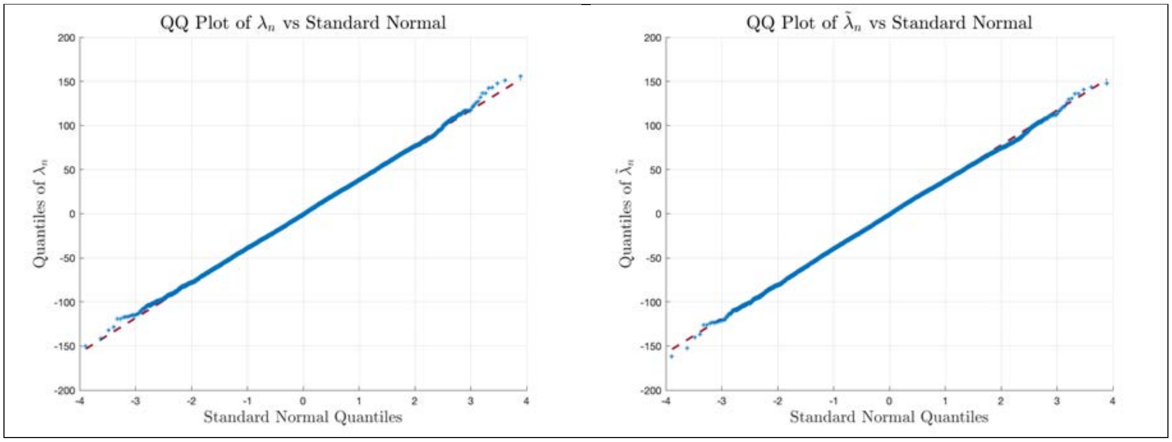

Figure 4.3. λ_T QQ plots from MatLab optimization software and Taylor approximation. $T = 70k$, p_i's and "2 low" q_i's from Tables (4.1) and (4.2).

4.1.2 MSE Reductions between Naive, ML, and CV

In this section we examine the reductions in Mean Squared Error of each estimator as T is increased. Of note "cn" indicates "censored naive", and "uc" indicates "uncensored". Further, each of these four separate trials were taken across 10k different realizations of each estimator for each of the values of T per trial with the p_i's from Table (4.1) and the q_i's from the "2 low" column in Table (4.2).

Table (4.4) demonstrates that as T increases, all of the estimators improve but it also shows that the difference between the mean squared error of the CV and ML estimators is very small, indicating that they are asymptotically identical. Further note that all four of the listed estimators have a variance that appears to decrease by a factor of ten every time T is increased by a factor of 10.

T	7,000	70,000	700,000	7,000,000
$\text{MSE}(\bar{Z}^{\text{cn}})$	1.3687e^{-1}	1.3592e^{-2}	1.3918e^{-3}	1.3714e^{-4}
$\text{MSE}(\bar{Z}^{\text{ml}})$	4.4335e^{-3}	4.2672e^{-4}	4.2957e^{-5}	4.2182e^{-6}
$\text{MSE}(\bar{Z}^{\text{cv}})$	4.4968e^{-3}	4.2733e^{-4}	4.2961e^{-5}	4.2182e^{-6}
$\text{MSE}(\bar{Z}^{\text{uc}})$	1.4195e^{-4}	1.4325e^{-5}	1.3947e^{-6}	1.4595e^{-7}

Table 4.4. Comparison of estimator MSEs as T increases from $7e^3$ to $7e^6$ with uniform censoring (q level "uniform" from Table (4.2)).

While more exploration can be conducted into the effect of adjusting the strata endpoints combined with changing the censoring scheme's probability mass function (namely, adjusting the q_i's), we only look at a few different q_i levels while keeping the fixed set of strata endpoints. Table (4.5) demonstrates that the effectiveness of the CV and ML estimators is connected to the q_i's. Of note, the $\text{MSE}^{\text{red}}_{\text{cn}\to\text{cv}}$ row in Table (4.5) shows that in this specific setting, the CV estimator results in a mean squared error reduction of roughly 99% in the "5 low" censoring scheme (the highest for this set of q_i's) and at worst a reduction of just above 95% in the "4 low" setting. These values are calculated by the following simple formula: $\text{MSE}^{\text{red}}_{\text{cn}\to\text{cv}} = 1 - \frac{\text{MSE}(\bar{Z}^{\text{cv}})}{\text{MSE}(\bar{Z}^{\text{cn}})}$.

$\boldsymbol{q}$ levels	1 low	2 low	3 low	4 low	5 low	6 low	uniform
MSE($\bar{Z}^{\text{cn}}$)	6.064e^{-3}	1.382e^{-2}	2.264e^{-2}	2.670e^{-2}	1.428e^{-2}	2.095e^{-2}	1.775e^{-3}
MSE($\bar{Z}^{\text{ml}}$)	2.368e^{-4}	4.456e^{-4}	8.454e^{-4}	1.183e^{-3}	1.577e^{-4}	4.095e^{-4}	7.838e^{-5}
MSE($\bar{Z}^{\text{cv}}$)	2.369e^{-4}	4.461e^{-4}	8.464e^{-4}	1.183e^{-3}	1.576e^{-4}	4.110e^{-4}	7.839e^{-5}
MSE($\bar{Z}^{\text{uc}}$)	1.487e^{-5}	1.470e^{-5}	1.327e^{-5}	1.381e^{-5}	1.391e^{-5}	1.395e^{-5}	1.395e^{-5}
$\text{MSE}^{\text{red}}_{\text{cn}\to\text{cv}}$	0.9609	0.9677	0.9626	0.9557	0.9890	0.9804	0.9558
$\text{MSE}^{\text{red}}_{\text{cn}\to\text{uc}}$	0.9975	0.9989	0.9994	0.9995	0.9990	0.9993	0.9921

Table 4.5. Estimator MSEs as the censoring algorithm varies with a T of 70k for each of the q_i levels defined in Table (4.2). Note, $\text{MSE}^{\text{red}}_{\text{cn}\to\text{cv}}$ denotes the reduction in estimator MSE of CV as compared to CN.

4.1.3 Distribution and Variance of ML and CV Estimators

In this section we examine the analytically derived variance reductions between the naive, ML, CV, and Uncensored estimators against their respective sample variances taken across 10^4 different realizations of each estimator for each of the levels of T and also across a range of μ values for the underlying Z distribution. Figures (4.4) and (4.5) show the histograms and QQ Plots for both the CV and ML estimators for the settings listed in their respective captions. What these figures show is that from a numerical perspective, these estimators do indeed appear to be normally distributed.

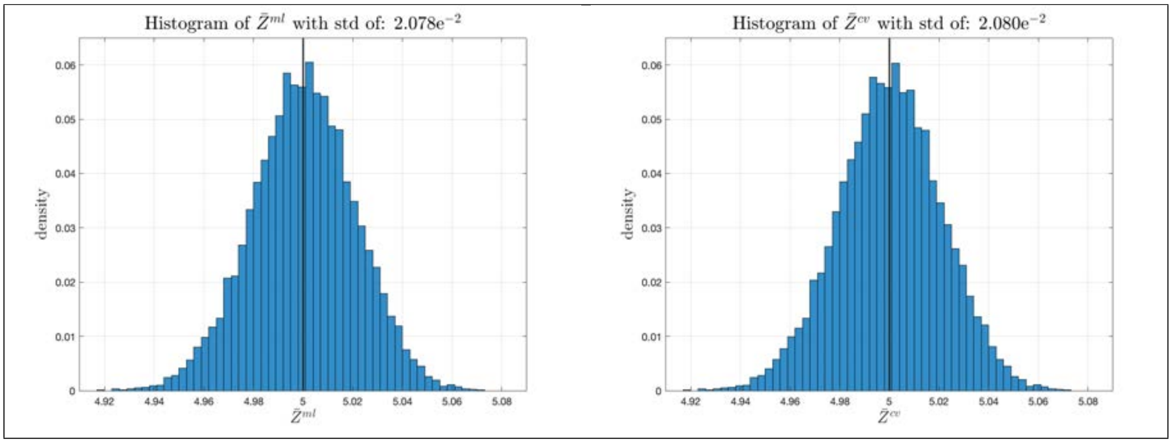

Figure 4.4. Histograms for $\bar{Z}^{ml}_T$ and $\bar{Z}^{cv}_T$ from MatLab simulation. $T = 70k$, p_i's and "2 low" q_i's from Tables (4.1) and (4.2) across 10^4 replications.

Of note, Table (4.6) demonstrates numerically not only that the differences between the variance of both the CV and ML estimators goes to zero asymptotically but also

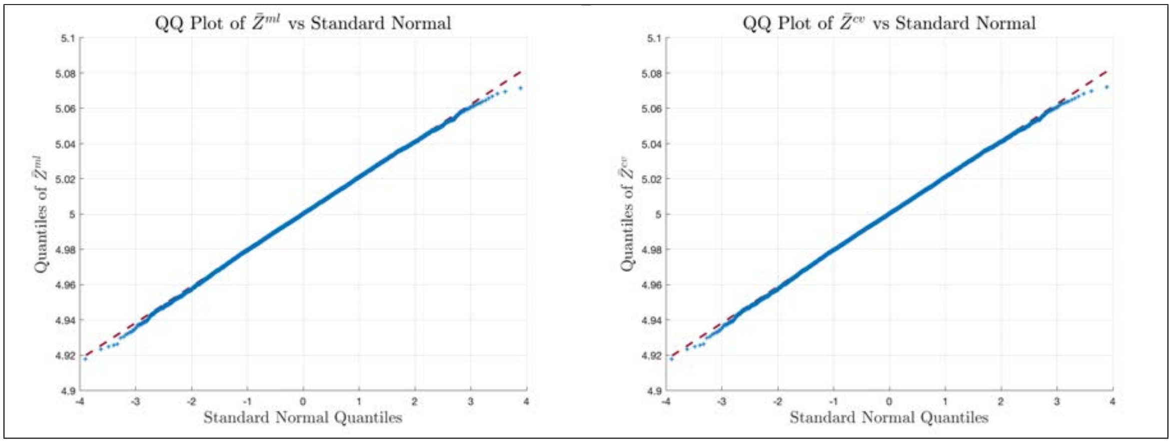

Figure 4.5. QQ Plots of $\bar{Z}_T^{ml}$ and $\bar{Z}_T^{cv}$ from MatLab simulation. $T = 70k$, p_i's and "2 low" q_i's from Tables (4.1) and (4.2) across 10^4 replications.

that the difference between the theoretical variances of the CV and ML estimators and the sample variances also goes to zero asymptotically. The listed theoretical variances are calculated based upon the primary results from Theorems 3.2.5 and 3.3.1. It also backs up the theoretical calculations in that for each of the estimators, the sample variances resulting from simulation match (within a couple percent) the theoretical results derived in Chapter 3, specifically, Theorems 3.2.5 and 3.3.1. Further, this table allows us to see the variance reduction from the "worst case" (censored naive), to the censored setting with CV and ML, and then finally to the "best case" (uncensored). Further, this table shows the variance reduction as the level of T increases. Specifically, an increase of $10T$ results in the CV and ML variances decreasing by $\frac{1}{10}$.

One very interesting result that was first observed in the numerical experiments can be seen in Table (4.7). Namely, that the censored naive estimator becomes more unstable as the mean of the underlying distribution Z is shifted or increased. Conversely, the CV and ML estimators are robust against that shift. Specifically, compare the $\mathrm{Var}(\bar{Z}^{\mathrm{cn}})$ and $\mathrm{Var}(\bar{Z}^{\mathrm{cv}})$ rows or the $S^2(\bar{Z}^{\mathrm{cn}})$, $S^2(\bar{Z}^{\mathrm{cv}})$, and $S^2(\bar{Z}^{\mathrm{ml}})$ rows against each other in Table (4.7). Of note, in Table (4.7), the strata are shifted with the mean of Z such that the p_i's remain constant across the various μ's.

T	7,000	70,000	700,000	7,000,000
$\mathrm{Var}(\bar{Z}^{\mathrm{cn}})$	5.9923e^{-2}	5.9923e^{-3}	5.9923e^{-4}	5.9923e^{-5}
$S^2(\bar{Z}^{\mathrm{cn}})$	6.0313e^{-2}	6.0810e^{-3}	6.0341e^{-4}	5.8696e^{-5}
$\mathrm{Var}(\bar{Z}^{\mathrm{cv}})$	2.5440e^{-3}	2.5440e^{-4}	2.5440e^{-5}	2.5440e^{-6}
$S^2(\bar{Z}^{\mathrm{cv}})$	2.5508e^{-3}	2.5722e^{-4}	2.5268e^{-5}	2.5764e^{-6}
$S^2(\bar{Z}^{\mathrm{ml}})$	2.5452e^{-3}	2.5717e^{-4}	2.5266e^{-5}	2.5764e^{-6}
$\mathrm{Var}(\bar{Z}^{\mathrm{uc}})$	1.4286e^{-4}	1.4286e^{-5}	1.4286e^{-6}	1.4286e^{-7}
$S^2(\bar{Z}^{\mathrm{uc}})$	1.4120e^{-4}	1.4491e^{-5}	1.4451e^{-6}	1.4219e^{-7}

Table 4.6. Sample and theoretical estimator variances for T varying from $7e^3$ to $7e^6$ with q_i's from the "1 low" column of Table (4.2) and 10^4 replications.

μ	0	2	5	10	25	50	200
$\mathrm{Var}(\bar{Z}^{\mathrm{cn}})$	4.003e^{-4}	1.925e^{-3}	1.383e^{-2}	6.010e^{-2}	3.970e^{-1}	1.619e^{+0}	2.628e^{+1}
$S^2(\bar{Z}^{\mathrm{cn}})$	3.982e^{-4}	1.926e^{-3}	1.403e^{-2}	6.107e^{-2}	3.975e^{-1}	1.602e^{+0}	2.651e^{+1}
$\mathrm{Var}(\bar{Z}^{\mathrm{cv}})$	2.010e^{-4}	4.284e^{-4}	4.292e^{-4}	4.286e^{-4}	4.286e^{-4}	4.286e^{-4}	4.286e^{-4}
$S^2(\bar{Z}^{\mathrm{cv}})$	1.950e^{-4}	4.227e^{-4}	4.311e^{-4}	4.243e^{-4}	4.278e^{-4}	4.242e^{-4}	4.339e^{-4}
$S^2(\bar{Z}^{\mathrm{ml}})$	1.941e^{-4}	4.222e^{-4}	4.305e^{-4}	4.237e^{-4}	4.272e^{-4}	4.237e^{-4}	4.333e^{-4}
$\mathrm{Var}(\bar{Z}^{\mathrm{uc}})$	1.429e^{-5}	1.429e^{-5}	1.429e^{-5}	1.429e^{-5}	1.429e^{-5}	1.429e^{-5}	1.429e^{-5}
$S^2(\bar{Z}^{\mathrm{uc}})$	1.416e^{-5}	1.410e^{-5}	1.448e^{-5}	1.429e^{-5}	1.422e^{-5}	1.431e^{-5}	1.409e^{-5}

Table 4.7. Sample and theoretical estimator variances as μ varies from 0 to 200 with q_i's from the "2 low" column of Table (4.2) with 10^4 replications.

4.2 Combinatorial Setting

In this setting we look at the case where instead of $|B_t| = 1$, this constraint is relaxed to a combination of cells such that $0 < |B_t| < m$. The last portion of the inequality is due to the fact that if $|B_t| = m$, we have the uncensored case. Further, if $|B_t| = 0$, the analyst is unable to observe Z_t at all. In this section, we again use a shifted standard normal, specifically, $Z \sim N(\mu = 5, \sigma = 1)$. But, to reduce overall simulation run times, we partition the support of Z into six ($m = 6$) cell intervals or strata using the following partition points of the real numbers: $\{4,\ 5,\ 6,\ 7,\ 8\}$. This set of partitions results in these cells: $c_1 = (-\infty, 4), c_2 = (4, 5), \ldots, c_6 = (8, \infty)$. This partition of Z, results in the specific values for the p_i's as displayed in Table (4.8).

Further, we use a multinomial distribution to guide the external censoring algorithm with a uniform probability mass across all possible combinations. This resulted in $q_i = \frac{31}{62} = \frac{1}{2} = P(c_i \in B_t)$ and $q_{ij} = \frac{15}{62} = P(c_i \in B_t, c_j \in B_t)$. Further, all results in this section come from 10^3 total simulations. Hence, the sample variances and Mean Squared Errors are taken across one thousand simulation results.

p_1	p_2	p_3	p_4	p_5	p_6
1.587e^{-1}	3.413e^{-1}	3.413e^{-1}	1.359e^{-1}	2.140e^{-2}	1.350e^{-3}

Table 4.8. p_i's for the shifted gaussian random variable $Z \sim N(\mu = 5, \sigma = 1)$ given $|B_t| > 1$.

4.2.1 MSE Reductions between Naive, ML, and CV

In this section, Table (4.9) examines the reductions in Mean Squared Error of each estimator as T is increased similar to the singleton section. Further, each of these four separate trials were taken across 10^3 different realizations of each estimator for each of the values of T per trial with the p_i's from Table (4.8).

T	600	6,000	60,000	600,000
MSE($\bar{Z}^{\text{cn}}$)	3.6055e^{-2}	3.3854e^{-3}	3.3789e^{-4}	3.4670e^{-5}
MSE($\bar{Z}^{\text{cv}}$)	3.2083e^{-3}	3.2056e^{-4}	2.9899e^{-5}	3.0878e^{-6}
MSE($\bar{Z}^{\text{ml}}$)	3.2063e^{-3}	3.2055e^{-4}	2.9899e^{-5}	3.0878e^{-6}
MSE($\bar{Z}^{\text{uc}}$)	1.6757e^{-3}	1.7663e^{-4}	1.6280e^{-5}	1.6187e^{-6}

Table 4.9. Estimator MSEs as T varies with uniform censoring.

Table (4.9) while demonstrating that as T increases, all of the estimators improve, also shows that the difference between the mean squared error of the CV and ML estimators is very small, indicating that they are asymptotically identical (confirming the results from Chapter 3). Further note that all four of the estimators have a MSE that appears to decrease by a factor of ten every time T is increased by a factor of 10. Of note, all estimator realizations (per simulation) are computed from the same simulation, namely $(Z_1, B_1), \ldots, (Z_t, B_t), \ldots, (Z_T, B_T)$. What this means, is that for the same data and sufficiently large T, the difference in estimation between CV and ML becomes vanishingly small.

4.2.2 Variance (theory vs. sample) of ML and CV Estimators

In this section we again examine the analytically (via Theorems 3.2.5 and 3.3.1) derived variance reductions between the naive, CV, and uncensored estimators against their respective sample variances taken across 10^3 different realizations of each estimator for each of the levels of T in the combinatorial setting. The results of this numerical analysis is provided in Table (4.10).

T	600	6,000	60,000	600,000
$\text{Var}(\bar{Z}^{\text{cn}})$	3.4376e^{-2}	3.4376e^{-3}	3.4376e^{-4}	3.4376e^{-5}
$S^2(\bar{Z}^{\text{cn}})$	3.6079e^{-2}	3.3883e^{-3}	3.3787e^{-4}	3.4704e^{-5}
$\text{Var}(\bar{Z}^{\text{cv}})$	3.0718e^{-3}	3.0718e^{-4}	3.0718e^{-5}	3.0718e^{-6}
$S^2(\bar{Z}^{\text{cv}})$	3.2100e^{-3}	3.2067e^{-4}	2.9874e^{-5}	3.0906e^{-6}
$S^2(\bar{Z}^{\text{ml}})$	3.2081e^{-3}	3.2066e^{-4}	2.9874e^{-5}	3.0906e^{-6}
$\text{Var}(\bar{Z}^{\text{uc}})$	1.6667e^{-3}	1.6667e^{-4}	1.6667e^{-5}	1.6667e^{-6}
$S^2(\bar{Z}^{\text{uc}})$	1.6769e^{-3}	1.7650e^{-4}	1.6249e^{-5}	1.6189e^{-6}

Table 4.10. Comparison between sample and theoretical estimator variances for T varying from $6e^2$ to $6e^5$ with 10^3 simulation replications.

Table (4.10) combined with the singleton results in Table (4.6) numerically confirm Theorems 3.2.5, 3.3.1, and 3.4.1 from Chapter 3. Namely, that in both the Singleton and Combinatorial settings, the difference between the variance of both the CV and ML estimators goes to zero asymptotically and that the difference between the theoretical variance of the CV and ML estimators and the sample variances also goes to zero asymptotically. Further, this table allows us to again see the variance reduction from the "worst case" (censored naive), to the censored setting with CV and ML, and then finally to the "best case" (uncensored). This table also shows the variance reduction as the level of T (number of die rolls, border crossings, or in general the number of samples) increases. Specifically, an increase from T to $10T$ results in the CV and ML variances decreasing by $\frac{1}{10}$, the same as the singleton setting.

CHAPTER 5:
Conclusion

The key advances or insights that this research provides are the connection between the methods of control variates and empirical or maximum likelihood in this censoring setting and their significant variance reduction (up to roughly 99% in some cases, see Table (4.5)) as compared to the censored naive estimator in the singleton setting using the exact same data. Further, both the CV and ML estimators are robust to a shift in the mean of the underlying distribution (see Table (4.7)) while the censored naive estimator becomes increasingly unstable (variance increases) for large means. Finally, the asymptotic connection between the CV and ML estimators enables the analyst to use the computationally inexpensive CV estimator with all of the desirable qualities of an MLE without its computational cost resulting from the optimization problem which requires solving a root equation inherent to MLE. Therefore, this research fills a gap in the literature by providing the asymptotic connection between the methods of control variates and maximum likelihood estimation within the censoring setting.

The following list provides some natural extensions to this dissertation that the author hopes to continue work on in the near future. This is not intended to be comprehensive.

Infinitely Many Strata

In this dissertation, m is constrained to be finite. An interesting extension to explore is relaxing this constraint to allow m to be countably infinite. This will require some modifications to the proof but will enable the results to be applied to a broader range of application settings.

Leveraging Additional Information about the p_i's

What if the analyst has additional information regarding Z? Specifically, for this dissertation the analyst only knows that $\sum_{i=1}^{m} p_i = 1$. But, what if the analyst also knew that say $p_i + p_j = 0.3$? This obviously should enable the estimators to perform even better resulting in even larger variance reductions over the censored naive estimator.

Leveraging Known p_i's and q_i's

Going a step beyond the last section, what if the analyst knew both the p_i's and the q_i's? This should result in an even larger variance reduction. Of note, known p_i's corresponds to the standard stratification methods examined in the literature. Also, knowing both enables the analyst to develop far more robust and strong estimators.

Intelligently Adjusting the Censoring Scheme

Further, what if the analyst could adjust the external censoring scheme? This additional constraint removal opens up numerous possibilities for more intelligently selecting which strata will be uncensored each time period.

Extension of Censoring Process to Continuous B

The idea here, is that instead of creating a discrete partition, $\mathcal{C}$ of the sample space Ω of the unknown random variable Z, we allow the "observed" interval or intervals to be continuous such that for $B \in \mathbb{R}$, B has a probability density function.

Bibliography

Akin EW (2017) *A multi-armed bandit approach to following a Markov Chain*. M.S. thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA, URL https://calhoun.nps.edu/handle/10945/55572.

Aldrich J (1997) RA Fisher and the making of maximum likelihood 1912-1922. *Statistical science* 12(3):162–176.

Asmussen S, Glynn PW (2007) *Stochastic simulation: Algorithms and analysis*, volume 57 (Springer Science & Business Media).

Banerjee M, et al. (2007) Likelihood based inference for monotone response models. *The Annals of Statistics* 35(3):931–956.

Bernoulli D (1766) Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Histoire de l'Acad., Roy. Sci. (Paris) avec Mem* 1–45.

Chen K, Zhou M (2003) Non-parametric hypothesis testing and confidence intervals with doubly censored data. *Lifetime Data Analysis* 9(1):71–91.

Cohen Jr AC (1950) Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *The Annals of Mathematical Statistics* 557–569.

Cox DR (2006) *Principles of statistical inference* (Cambridge university press).

Cramér H (2016) *Mathematical Methods of Statistics (PMS-9)*, volume 9 (Princeton, NJ: Princeton University Press).

Davarzani N, Parsian A (2011) Statistical inference for discrete middle-censored data. *Journal of Statistical Planning and Inference* 141(4):1455–1462.

Devore JL (2015) *Probability and Statistics for Engineering and the Sciences* (Boston, MA: Cengage Learning).

Durrett R (2019) *Probability: Theory and Examples*, volume 49 (Cambridge University Press), 5th edition.

Edgeworth FY (1908*a*) On the probable errors of frequency-constants. *Journal of the Royal Statistical Society* 71(2):381–397.

Edgeworth FY (1908*b*) On the probable errors of frequency-constants (contd.). *Journal of the Royal Statistical Society* 71(4):651–678.

Fisher R (1931) Estimation of mean and standard deviation of normal population (truncated). *Mathematical Tables* 1:33–34.

Fygenson M, Zhou M, et al. (1994) On using stratification in the analysis of linear regression models with right censoring. *The Annals of Statistics* 22(2):747–762.

Gill RD, Wellner JA, Præstgaard J (1989) Non-and semi-parametric maximum likelihood estimators and the von Mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics* 97–128.

Glasserman P, Yu B (2005) Large sample properties of weighted Monte Carlo estimators. *Operations Research* 53(2):298–312.

Glynn PW, Szechtman R (2002) Some new perspectives on the method of control variates. *Monte Carlo and Quasi-Monte Carlo Methods 2000*, 27–49 (Springer).

Gupta A (1952) Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika* 39(3/4):260–273.

Hald A (1949) Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Scandinavian Actuarial Journal* 1949(1):119–134.

Jammalamadaka SR, Mangalam V (2003) Nonparametric estimation for middle-censored data. *Journal of nonparametric statistics* 15(2):253–265.

Kharroubi SA (2018) Posterior simulation via the exponentially tilted signed root log-likelihood ratio. *Computational Statistics* 33(1):213–234.

Koul H, Susarla Vv, Van Ryzin J, et al. (1981) Regression analysis with randomly right-censored data. *The Annals of statistics* 9(6):1276–1288.

Law AM (2007) *Simulation Modeling and Analysis* (McGraw-Hill), 4th edition.

Lehmann EL, Casella G (1998) *Theory of Point Estimation* (Boston, MA: Springer Science & Business Media), ISBN 0-387-98502-6.

Leurgans S (1987) Linear models, random censoring and synthetic data. *Biometrika* 74(2):301–309.

Murphy SA, van der Vaart AW, et al. (1997) Semiparametric likelihood ratio inference. *The Annals of Statistics* 25(4):1471–1509.

Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231(694-706):289–337.

Owen AB (2001) *Empirical Likelihood* (CRC press).

Pearson K, Lee A (1908) On the generalised probable error in multiple normal correlation. *Biometrika* 6(1):59–68.

Pfanzagl J (2011) *Parametric statistical theory* (Walter de Gruyter).

Rao CR (1992) Information and the accuracy attainable in the estimation of statistical parameters. *Breakthroughs in statistics*, 235–247 (Springer).

Ross SM (1972) *Introduction to Probability Models* (Academic Press), 1st edition.

Ross SM (2014) *Introduction to Probability Models* (Academic press), 11th edition.

Sargan JD (1976) Econometric estimators and the Edgeworth approximation. *Econometrica: Journal of the Econometric Society* 421–448.

Stevens W (1937) The truncated normal distribution. *Annals of Applied Biology* 24(4):815–852.

Szechtman R, Glynn PW (2001) Constrained Monte Carlo and the method of control variates. *Proceeding of the 2001 Winter Simulation Conference (Cat. No. 01CH37304)*, volume 1, 394–400 (IEEE).

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics* 9(1):60–62.

Zheng Z, Glynn PW (2017) A CLT for infinitely stratified estimators, with applications to debiased MLMC. *ESAIM: Proceedings and Surveys* 59:104–114.

Zhou M (2005) Empirical likelihood ratio with arbitrarily censored/truncated data by EM algorithm. *Journal of Computational and Graphical Statistics* 14(3):643–656.

Zivot E, Wang J (2007) *Generalized Method of Moments*, volume 191, 785–845 (New York, NY: Springer Science & Business Media), ISBN 978-0-387-32348-0, URL http://dx.doi.org/10.1007/978-0-387-32348-0_21.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California